21世纪高等学校规划教材 | 计算机科学与技术



路由和交换技术

沈鑫剡 编著

清华大学出版社

21 世纪高等学校规划教材·计算机科学与技术

路由和交换技术

沈鑫剡　编著

清华大学出版社
北　京

内 容 简 介

本教材详细讨论了 MAC 帧和 IP 分组端到端传输过程中涉及的设备、协议和算法。具体内容包括以太网交换机结构、生成树算法、链路聚合算法、VLAN 划分、路由器结构、路由协议、组播、网络地址转换、三层交换和 IPv6 网络等。

本教材在具体网络环境下深入讨论交换式以太网和互联网络的基本原理、算法、协议及各协议间的相互作用过程，既有理论总结，又有应用实例。结合当前主流厂家的交换机和路由器设备，向读者介绍完整、深入的路由和交换技术，解决了其他教材中存在的内容和实际应用脱节的问题，使读者能够学以致用。

本教材以通俗易懂、循序渐进的方式叙述路由和交换技术，并通过大量的例子来加深读者对路由和交换技术的理解，是一本理想的计算机网络工程专业的路由和交换技术教材，对从事校园网、企业网设计和实施的工程技术人员和从事交换机、路由器研发的科研人员，也是一本非常好的参考书。

图书在版编目（CIP）数据

路由和交换技术/沈鑫剡编著. —北京：清华大学出版社，2013.2
（21 世纪高等学校规划教材 · 计算机科学与技术）
ISBN 978-7-302-29842-7

Ⅰ. ①路… Ⅱ. ①沈… Ⅲ. ①计算机网络－路由选择－高等学校－教材 ②计算机网络－信息交换机－高等学校－教材 Ⅳ. ①TN915.05

中国版本图书馆 CIP 数据核字(2012)第 197273 号

责任编辑：刘向威 薛 阳
封面设计：
责任校对：梁 毅
责任印制：

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 刷 者：
装 订 者：
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：21.25　　**字　　数**：532 千字
版　　次：2013 年 2 月第 1 版　　**印　　次**：2013 年 2 月第1 次印刷
印　　数：1～ 000
定　　价： .00 元

产品编号：046711-01

编审委员会成员

（按地区排序）

出版说明

随着我国改革开放的进一步深化，高等教育也得到了快速发展，各地高校紧密结合地方经济建设发展需要，科学运用市场调节机制，加大了使用信息科学等现代科学技术提升、改造传统学科专业的投入力度，通过教育改革合理调整和配置了教育资源，优化了传统学科专业，积极为地方经济建设输送人才，为我国经济社会的快速、健康和可持续发展以及高等教育自身的改革发展做出了巨大贡献。但是，高等教育质量还需要进一步提高以适应经济社会发展的需要，不少高校的专业设置和结构不尽合理，教师队伍整体素质亟待提高，人才培养模式、教学内容和方法需要进一步转变，学生的实践能力和创新精神亟待加强。

教育部一直十分重视高等教育质量工作。2007 年 1 月，教育部下发了《关于实施高等学校本科教学质量与教学改革工程的意见》，计划实施“高等学校本科教学质量与教学改革工程”（简称“质量工程”），通过专业结构调整、课程教材建设、实践教学改革、教学团队建设等多项内容，进一步深化高等学校教学改革，提高人才培养的能力和水平，更好地满足经济社会发展对高素质人才的需要。在贯彻和落实教育部“质量工程”的过程中，各地高校发挥师资力量强、办学经验丰富、教学资源充裕等优势，对其特色专业及特色课程（群）加以规划、整理和总结，更新教学内容、改革课程体系，建设了一大批内容新、体系新、方法新、手段新的特色课程。在此基础上，经教育部相关教学指导委员会专家的指导和建议，清华大学出版社在多个领域精选各高校的特色课程，分别规划出版系列教材，以配合“质量工程”的实施，满足各高校教学质量和教学改革的需要。

为了深入贯彻落实教育部《关于加强高等学校本科教学工作，提高教学质量的若干意见》精神，紧密配合教育部已经启动的“高等学校教学质量与教学改革工程精品课程建设工作”，在有关专家、教授的倡议和有关部门的大力支持下，我们组织并成立了“清华大学出版社教材编审委员会”（以下简称“编委会”），旨在配合教育部制定精品课程教材的出版规划，讨论并实施精品课程教材的编写与出版工作。“编委会”成员皆来自全国各类高等学校教学与科研第一线的骨干教师，其中许多教师为各校相关院、系主管教学的院长或系主任。

按照教育部的要求，“编委会”一致认为，精品课程的建设工作从开始就要坚持高标准、严要求，处于一个比较高的起点上。精品课程教材应该能够反映各高校教学改革与课程建设的需要，要有特色风格、有创新性（新体系、新内容、新手段、新思路，教材的内容体系有较高的科学创新、技术创新和理念创新的含量）、先进性（对原有的学科体系有实质性的改革和发展，顺应并符合 21 世纪教学发展的规律，代表并引领课程发展的趋势和方向）、示范性（教材所体现的课程体系具有较广泛的辐射性和示范性）和一定的前瞻性。教材由个人申报或各校推荐（通过所在高校的“编委会”成员推荐），经“编委会”认真评审，最后由清华大学出版

社审定出版。

目前，针对计算机类和电子信息类相关专业成立了两个“编委会”，即“清华大学出版社计算机教材编审委员会”和“清华大学出版社电子信息教材编审委员会”。推出的特色精品教材包括：

（1）21世纪高等学校规划教材·计算机应用——高等学校各类专业，特别是非计算机专业的计算机应用类教材。

（2）21世纪高等学校规划教材·计算机科学与技术——高等学校计算机相关专业的教材。

（3）21世纪高等学校规划教材·电子信息——高等学校电子信息相关专业的教材。

（4）21世纪高等学校规划教材·软件工程——高等学校软件工程相关专业的教材。

（5）21世纪高等学校规划教材·信息管理与信息系统。

（6）21世纪高等学校规划教材·财经管理与应用。

（7）21世纪高等学校规划教材·电子商务。

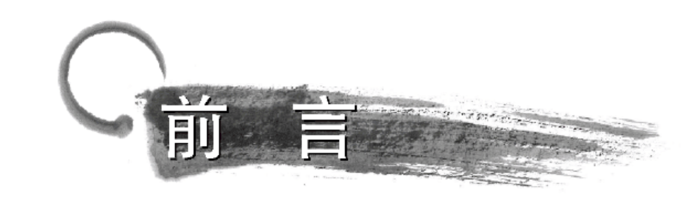
（8）21世纪高等学校规划教材·物联网。

清华大学出版社经过三十多年的努力，在教材尤其是计算机和电子信息类专业教材出版方面树立了权威品牌，为我国的高等教育事业做出了重要贡献。清华版教材形成了技术准确、内容严谨的独特风格，这种风格将延续并反映在特色精品教材的建设中。

清华大学出版社教材编审委员会

联系人：魏江江

E-mail：weijj@tup.tsinghua.edu.cn



前言

"路由和交换技术"课程的教学目标有三：一是使学生具备设计、实施校园网和企业网的工程能力；二是具备研发交换机和路由器的能力；三是具备 MAC 帧和 IP 分组端到端传输过程所涉及的算法和协议的分析、设计和实现能力。但目前市场上的《路由和交换技术》教材主要可以分为两类，一类教材的内容与《计算机网络》教材内容高度重叠，对于交换机和路由器结构，交换式以太网和互联网相关算法和协议的工作原理、实现过程，涉及较少，因而无法培养学生研发交换机和路由器、实施网络工程的能力。另一类教材像是交换机和路由器配置指南，主要讨论常见交换机和路由器设备的配置过程，对于交换机和路由器结构，交换式以太网和互联网相关算法和协议的工作原理、实现过程，仍然涉及较少，虽然可以使学生具有一定设计、实施校园网和企业网的能力，但无法使学生具有研发交换机和路由器的能力与相关算法和协议的分析、设计和实现能力。

本教材的特点：一是详细讨论交换机和路由器结构，交换式以太网和互联网相关算法和协议的工作原理、实现过程，提供完成校园网、企业网方案设计和实施所需要的交换式以太网和互联网的知识。二是在具体网络环境下深入讨论交换式以太网和互联网相关算法和协议的工作原理及各协议间的相互作用过程，为学生提供透彻、完整的交换式以太网和互联网知识。三是结合主流厂家设备讨论交换和路由技术，并将它们讲深讲透，让读者能够学以致用。四是通过大量例题解析为读者提供运用所学知识分析、解决问题的方法和步骤。五是通过对大量取自实际应用的案例的分析，为读者提供设计、实施校园网和企业网与分析、设计和实现相关算法和协议的思路。

"路由和交换技术"课程是一门实验性很强的课程，掌握交换机和路由器配置过程及交换式以太网和互联网设计、实施过程对于深入了解交换式以太网和互联网相关算法和协议的工作原理、实现过程非常有用。鉴于目前很少有学校可以提供能够完成各种规模校园网和企业网设计、实施实验的网络实验室，为此编写了作为指导学生利用 Cisco Packet Tracer 软件实验平台完成各种规模校园网和企业网设计、实施实验的实验指导书的配套教材《路由和交换技术实验及实训》。Cisco Packet Tracer 软件实验平台的人机界面非常接近实际配置过程，学生通过 Cisco Packet Tracer 软件实验平台可以完成教材内容涵盖的全部实验，建立与现实网络世界相似的应用环境，真正掌握基于 Cisco 设备完成交换式以太网和互联网设计、配置和调试的方法和步骤。

作为一本无论在内容组织、叙述方法还是教学目标都和已有《路由和交换技术》教材有一定区别的新教材，错误和不足之处在所难免，殷切希望使用该教材的老师和学生批评指正，也殷切希望读者能够就教材内容和叙述方式提出宝贵建议和意见，以便进一步完善教材内容。作者 E-mail 地址为：shenxinshan@163.com。

编　者

2012 年 5 月于南京

目 录

第 1 章　交换机和交换式以太网 …… 1

1.1　以太网概述 …… 1
1.1.1　以太网发展过程 …… 1
1.1.2　以太网体系结构 …… 3
1.1.3　以太网拓扑结构 …… 4
1.2　以太网从共享到交换 …… 6
1.2.1　总线型以太网 …… 6
1.2.2　透明网桥与冲突域分割 …… 17
1.3　交换机转发方式和交换机结构 …… 24
1.3.1　交换机转发方式 …… 24
1.3.2　交换机结构 …… 25
1.4　以太网标准 …… 31
1.4.1　10Mb/s 以太网标准 …… 31
1.4.2　100Mb/s 以太网标准 …… 31
1.4.3　1Gb/s 以太网标准 …… 32
1.4.4　10Gb/s 以太网标准 …… 33
习题 …… 33

第 2 章　虚拟局域网 …… 38

2.1　广播域和广播传输方式 …… 38
2.1.1　单播传输方式和广播传输方式 …… 38
2.1.2　广播域 …… 39
2.1.3　传统分割广播域的方式 …… 40
2.2　VLAN 定义和分类 …… 40
2.2.1　VLAN 定义 …… 40
2.2.2　VLAN 分类 …… 41
2.3　基于端口划分 VLAN …… 44
2.3.1　单交换机 VLAN 划分过程 …… 44
2.3.2　跨交换机 VLAN 划分过程 …… 45
2.3.3　802.1Q 与 VLAN 内数据传输 …… 45
2.3.4　端口确定 MAC 帧所属 VLAN 规则 …… 47
2.3.5　VLAN 例题解析 …… 47

2.4 Cisco 基于 MAC 地址划分 VLAN 技术 …… 52
2.5 专用 VLAN …… 53
2.5.1 专用 VLAN 的作用 …… 53
2.5.2 Cisco 专用 VLAN 工作原理 …… 54
2.6 VLAN 属性注册协议 …… 59
2.6.1 GVRP 作用 …… 59
2.6.2 GARP …… 60
2.6.3 GVRP 工作原理 …… 62
2.6.4 VTP …… 65
2.6.5 GVRP 例题解析 …… 72
习题 …… 74

第 3 章 生成树协议 …… 78

3.1 生成树协议的作用 …… 78
3.1.1 环路引发广播风暴 …… 78
3.1.2 树型网络的弱可靠性 …… 79
3.1.3 生成树协议的由来和发展 …… 79
3.2 生成树协议工作过程 …… 80
3.2.1 生成树协议操作步骤 …… 80
3.2.2 生成树协议构建生成树过程 …… 81
3.2.3 生成树协议的容错功能 …… 85
3.2.4 端口状态和定时器 …… 86
3.2.5 网桥转发表刷新机制 …… 87
3.2.6 STP 例题解析 …… 90
3.3 快速生成树协议 …… 91
3.3.1 STP 的缺陷 …… 91
3.3.2 端口角色和端口状态 …… 92
3.3.3 端口状态快速迁移过程 …… 93
3.3.4 网桥转发表刷新机制 …… 95
3.3.5 RSTP 例题解析 …… 96
3.4 多生成树协议 …… 96
3.4.1 MSTP 的必要性 …… 96
3.4.2 MSTP 基本思想 …… 97
3.4.3 MSTP 工作过程 …… 98
习题 …… 103

第 4 章 以太网链路聚合 …… 105

4.1 链路聚合基础 …… 105
4.1.1 链路聚合含义 …… 105

4.1.2 链路聚合方式…… 106
4.1.3 端口属性…… 106
4.2 链路聚合机制 …… 106
4.2.1 功能组成…… 106
4.2.2 交换机通过聚合组转发 MAC 帧过程 …… 109
4.2.3 链路聚合组生成过程…… 110
4.3 链路聚合控制协议 …… 111
4.3.1 LACP 简介 …… 111
4.3.2 LACP 报文格式…… 112
4.3.3 LACP 工作过程 …… 112
习题…… 114

第 5 章 路由器和网络互连…… 116

5.1 网络互连 …… 116
5.1.1 网络互连需要解决的问题…… 116
5.1.2 信件投递过程的启示…… 117
5.1.3 端到端传输思路…… 118
5.1.4 IP 实现网络互连机制 …… 119
5.1.5 数据报 IP 分组交换网络 …… 120
5.1.6 路由器结构…… 122
5.2 网际协议 …… 123
5.2.1 IP 地址分类 …… 123
5.2.2 IP 地址分层分类的原因和缺陷 …… 125
5.2.3 无分类编址…… 129
5.2.4 IP 分组格式 …… 136
5.3 路由表和 IP 分组端到端传输过程…… 140
5.3.1 路由表建立过程…… 140
5.3.2 IP 分组端到端传输过程 …… 144
5.3.3 ARP 和地址解析过程 …… 146
5.4 虚拟路由器冗余协议 …… 148
5.4.1 容错网络结构…… 148
5.4.2 VRRP 工作原理 …… 149
5.4.3 VRRP 应用实例 …… 155
习题…… 156

第 6 章 路由协议…… 161

6.1 路由项分类 …… 161
6.1.1 直连路由项…… 161
6.1.2 静态路由项…… 162

6.1.3 动态路由项……163
6.1.4 静态路由项缺陷……164
6.2 路由协议基础……164
6.2.1 路由协议分类……165
6.2.2 路由协议要求……166
6.2.3 距离向量路由协议……167
6.2.4 链路状态路由协议……170
6.3 RIP……174
6.3.1 RIP 消息格式……174
6.3.2 RIP 工作过程……175
6.3.3 RIP 建立路由表实例……177
6.3.4 RIP 动态适应网络变化的过程……182
6.3.5 计数到无穷大和水平分割……183
6.4 OSPF……185
6.4.1 路由器确定自身链路状态……186
6.4.2 泛洪链路状态通告……193
6.4.3 构建路由表算法……195
6.4.4 OSPF 动态适应网络变化的过程……198
6.4.5 OSPF 和 RIP 的区别……199
6.4.6 OSPF 分区域建立路由表的过程……199
6.5 BGP……205
6.5.1 分层路由的原因……205
6.5.2 BGP 报文类型……206
6.5.3 BGP 工作机制……206
习题……210

第 7 章 组播……212

7.1 组播基本概念……212
7.1.1 组播与单播和广播的区别……212
7.1.2 组播地址……213
7.1.3 组播实现技术……214
7.2 IGMP……217
7.2.1 IGMP 消息类型和格式……217
7.2.2 IGMP 操作过程……218
7.2.3 IGMP 侦听……220
7.3 组播路由协议……223
7.3.1 DVMRP……223
7.3.2 PIM-SM……234
习题……242

第 8 章 网络地址转换 …… 244

8.1 NAT 基本概念 …… 244
8.1.1 NAT 定义 …… 244
8.1.2 私有地址空间 …… 245
8.1.3 NAT 应用 …… 246
8.1.4 NAT 引发的问题 …… 248
8.2 NAT 工作过程 …… 249
8.2.1 NAT 分类 …… 249
8.2.2 PAT …… 250
8.2.3 NAT …… 252
8.2.4 应用层网关 …… 254
8.3 NAT 应用方式 …… 255
8.3.1 双穴网络结构 …… 255
8.3.2 实现内部网络和外部网络通信 …… 257
8.3.3 实现内部网络之间通信 …… 259
8.3.4 解决内部网络与外部网络地址重叠问题 …… 261
习题 …… 264

第 9 章 三层交换机和三层交换 …… 266

9.1 三层交换机基础 …… 266
9.1.1 三层交换机产生背景 …… 266
9.1.2 三层交换机与路由器的区别 …… 271
9.1.3 校园网和三层交换机 …… 272
9.1.4 VLAN 互连实例 …… 274
9.2 三层交换过程 …… 277
9.2.1 三层交换机结构 …… 277
9.2.2 二层交换过程 …… 278
9.2.3 三层路由过程 …… 280
9.3 三层交换机应用方式 …… 282
9.3.1 IP 接口集中到单个三层交换机 …… 282
9.3.2 两个三层交换机同时定义所有 VLAN 对应的 IP 接口 …… 284
9.3.3 两个三层交换机分别定义两个 VLAN 对应的 IP 接口 …… 286
习题 …… 289

第 10 章 IPv6 …… 291

10.1 IPv4 的缺陷 …… 291
10.1.1 地址短缺问题 …… 291
10.1.2 复杂的分组首部 …… 292

10.1.3 QoS 实现困难 …… 292
10.1.4 安全机制先天不足 …… 292
10.2 IPv6 首部结构 …… 293
10.2.1 IPv6 基本首部 …… 293
10.2.2 IPv6 扩展首部 …… 295
10.3 IPv6 地址结构 …… 297
10.3.1 IPv6 地址表示方式 …… 297
10.3.2 IPv6 地址分类 …… 299
10.4 IPv6 操作过程 …… 303
10.4.1 邻站发现协议 …… 303
10.4.2 路由器建立路由表过程 …… 306
10.5 IPv6 over 以太网 …… 308
10.5.1 地址解析过程 …… 308
10.5.2 IPv6 组播地址和 MAC 组地址之间的关系 …… 310
10.5.3 IPv6 分组传输过程 …… 310
10.6 IPv6 网络和 IPv4 网络互连 …… 311
10.6.1 双协议栈技术 …… 311
10.6.2 隧道技术 …… 312
10.6.3 网络地址和协议转换技术 …… 314
习题 …… 320

英文缩写词 …… 323

参考文献 …… 326



第1章 交换机和交换式以太网

从共享式以太网发展到交换式以太网是以太网发展过程中的一次革命，交换机和以交换机为核心设备的交换式以太网的出现，使得交换成为 MAC 帧端到端传输机制的代名词。交换包含的内容非常广泛，MAC 帧端到端传输过程所涉及的算法和协议的实现机制都属于交换的范畴。交换式以太网发展过程，以太网终端之间传输路径建立过程和交换机 MAC 帧转发过程属于交换的基础知识部分。

1.1 以太网概述

1.1.1 以太网发展过程

1972 年底，Bob Metcalfe 和 David Boggs 设计了一套用于实现不同的 ALTO 计算机之间连接的网络。由于该网络是以 ALOHA 系统为基础的，且又连接了众多的 ALTO 计算机，Metcalfe 把该网络命名为 ALTO ALOHA 网络。ALTO ALOHA 网络于 1973 年 5 月 22 日首次开始运行。就在这一天，Metcalfe 将该网络改名为以太网，以此用于说明设计该网络的灵感来自于“电磁辐射可以通过发光的以太来传播”这一想法。

20 世纪 70 年代末，已经涌现出数十种局域网技术，以太网能够脱颖而出，登上局域网宝座的根本原因是 Metcalfe 版的以太网成为了产业标准。

多种原因导致 Xerox、Dec 和 Intel 联合起来开发以太网产品。三家联合的优势是显而易见的：Xerox 提供技术，Dec 有雄厚的技术力量，而且是以太网硬件的强有力的供应商，Intel 提供以太网硅片构件。1979 年 9 月 30 日，Dec、Intel 和 Xerox 公布了第三稿的“以太网，一种局域网——数据链路层和物理层规范 1.0 版”，这就是著名的以太网蓝皮书，也称为 DIX 版以太网 1.0 规范。DIX 版以太网 1.0 规范开始规定传输速率为 20Mb/s，最后降为 10Mb/s。

在 DIX 开展以太网标准化工作的同时，世界性专业组织 IEEE 组成一个定义与促进工业局域网(Local Area Network，LAN)标准的委员会，该委员会名叫 802 工程，以制定实现办公室环境下计算机连接的 LAN 标准为主要工作目标。1981 年 6 月，IEEE 802 工程决定组成 802.3 分委员会，以产生基于 DIX 工作成果的国际标准。1982 年 12 月 19 日，19 个公司宣布了新的 IEEE 802.3 草稿标准。1983 年该草稿最终以 IEEE 10BASE5 而面世。

1979 年 6 月，Bob Metcalfe 等人组建了计算机、通信和兼容性公司，即著名的 3Com

公司。

1980 年 8 月,3Com 公司宣布了它的第一个产品,用于 UNIX 的商业版 TCP/IP,该产品在 1980 年 12 月正式上市。1981 年 3 月,3Com 将第一批符合 802 标准的产品 3C100 收发器投放市场。1981 年底,3Com 公司开始销售 DEC PDP/11 系列和 VAX 系列的收发器和插卡,同时也销售在 Intel Multibus 和 Sun 微系统公司机器上使用的收发器和插卡。

1982 年 9 月 29 日,第一块为个人计算机(Personal Computer,PC)开发的 EtherLink 投放市场,并随卡提供相应的 DOS 驱动器软件。第一块 EtherLink 在以下多个方面取得突破。

(1) EtherLink 成为第一块在 IBM PC ISA 总线上使用的以太网适配器,这是以太网发展史上的一个里程碑。

(2) EtherLink 网络接口卡通过硅半导体集成工艺实现,它是第一块包含以太网 VLSI 控制器硅片的网络接口卡(Network Interface Card,NIC),由于硅片价格低,3Com 公司的 EtherLink 的价格比其他的网络接口卡和以前销售的收发器要便宜很多。

(3) 因为采用超大规模集成电路芯片节省了大量空间,EtherLink 适配器可以将收发器集成在网络接口卡上,省去了外接的 MAU 收发器。

随着个人计算机迅速占领市场,个人计算机联网的要求也日益迫切,EtherLink 生意火暴。1983 年,3Com 、ICL、HP 将细缆以太网的概念提交给 IEEE,不久 IEEE 就公布了细缆以太网的官方标准 10BASE2。

1986 年,SynOptics 开始进行在 UTP 电话线上运行 10Mb/s 以太网的研究工作,名叫 LATTIS NET 的第一个 SynOptics 产品于 1987 年 8 月 17 日正式投放市场,也就是在同一天,IEEE 802.3 工作组开始讨论在 UTP 上实现 10Mb/s 以太网的最佳方法,并在后来成为非屏蔽双绞线的官方标准 10BASET。10BASET 的出现,导致了结构化布线系统的兴起和发展。

传统共享介质以太网的缺陷是明显的,当网上用户数增多,总线负载加重,就会导致冲突频繁发生,使总线利用率急剧下降。为了解决这一问题,将以太网分段,每段以太网是一个独立的冲突域,多个不同的冲突域可以同时实现冲突域内终端之间的通信。为了实现连接在不同冲突域的两个终端之间通信,开发出一种叫网桥的产品,用网桥实现冲突域互连,以此实现连接在一个冲突域上的终端和连接在另一个冲突域上的终端之间的通信,由于网桥互连的多个冲突域可以同时实现冲突域内终端之间通信,网络的整体带宽得到提高。20 世纪 80 年代末,一种新型网桥——智能型多端口网桥开始出现。1990 年,一个完全不同的网桥——Kalpana Ether Switch EPS-700 面世。Ether Switch 具有下述功能特点。

(1) Ether Switch 和电话交换机相似,能够同时提供多条数据传输路径,使整体吞吐量得到显著提高。

(2) Ether Switch 使用一种名为"直通(cut-through)"的新桥接技术,其转发延迟比传统网桥使用的存储转发技术降低了一个数量级。

(3) Ether Switch 的推销员指出 Ether Switch 是网络交换器,而不是普通网桥,由此开辟了一个新的市场领域——网络交换机。

由于所有终端共享单条总线,共享介质以太网只能以半双工方式工作,终端在同一时间要么发送数据,要么接收数据,而不能同时发送和接收数据。网络交换机的出现,允许每一

端口只和一个终端传输数据，使得交换机端口和终端之间同时发送、接收数据成为可能，由此产生了以太网全双工通信标准，它使传输速度提高了一倍。

网络交换设备虽然是降低网络通信拥挤的最佳设备，但每个以太网交换机端口只能提供 10Mb/s 的通信流量，对于要求 10Mb/s 以上通信流量的应用，当时只能采用光纤分布式数据接口（Fiber Distributed Data Interface，FDDI），它是一个基于 100Mb/s 光纤的 LAN，极其昂贵。

1992 年下半年，新成立的 Grand Junction 公司开始研制 100Mb/s 以太网。对100Mb/s 以太网出现了两种技术方案：一种是继续保留现行以太网协议，一种是采用全新的 MAC 协议。前者方案得到了极大多数产品厂家的支持，这些厂家在 IEEE 802.3 工作组尚未做出决定之前，成立了快速以太网联盟（FEA），公布了它的 100BASE-TX 标准，并推出了第一台符合标准的集线器和网络接口卡（Network Interface Card，NIC）。

1995 年 3 月，IEEE 802.3u 标准获得通过，宣布快速以太网的时代来临。

1996 年 3 月，IEEE 组建了新的 802.3Z 工作组，负责研究吉比特以太网（俗称千兆以太网），并制订相应标准。很快，一些原来快速以太网的支持者和某些新的发起者组成了吉比特以太网联盟（GEA）。

到 1997 年底，3Com 公司已开始推出符合 802.3Z 标准草案的全套吉比特以太网设备，从吉比特交换机，快速以太网交换机的吉比特升级模块到吉比特以太网卡。1998 年 3 月，IEEE 802.3Z 标准获得通过。由于受冲突窗口的限制，吉比特以太网最好以全双工方式进行通信，否则通信距离将受到限制。

2002 年 7 月，IEEE 通过了 802.3ae 标准，开始了 10Gb/s 以太网（俗称万兆以太网）时代。

1.1.2　以太网体系结构

以太网标准的制定过程存在两条主线，一条主线是 Dec、Intel 和 Xerox 这三家公司在 1980 年 9 月制定并发表的关于以太网规约的第一个版本——DIX V1（DIX 由这三家公司名称的第一个字母组合而成），和在 1982 年修改发表的第二个版本——DIX Ethernet V2。另一条主线是电子和电气工程师协会（IEEE）802 委员会在 DIX Ethernet V2 基础上制定的第一个局域网标准，编号为 802.3。实际上，802.3 标准和 DIX Ethernet V2 还是有点差别的，但目前人们已习惯将符合 802.3 标准的局域网称做以太网。以太网并不是 IEEE 802 委员会制定的唯一局域网标准，在制定 802.3 标准以后，又陆续制定了多个不同的局域网标准，如令牌环网，由于不同局域网的链路层标准并不相同，为了给网络层提供统一的局域网功能界面，802 委员会将局域网的链路层分成两个子层：逻辑链路控制（Logical Link Control，LLC）子层和媒体接入控制（Medium Access Control，MAC）子层。因此，可以得出如图 1.1 所示的以以太网为传输网络的 TCP/IP 体系结构。

不同局域网的 MAC 子层是不同的，但 LLC 子层和网际层之间的接口界面是相同的，也就是说 LLC 子层屏蔽了由于多种局域网并存而造成的 MAC 子层的不同，就像 BIOS 屏蔽了主板的差异一样。

以太网的物理层主要解决和传输媒体之间的接口，数字信号 0、1 的表示方式，数字信号的同步等问题。MAC 子层解决和以太网传输数据有关的其他一些问题。

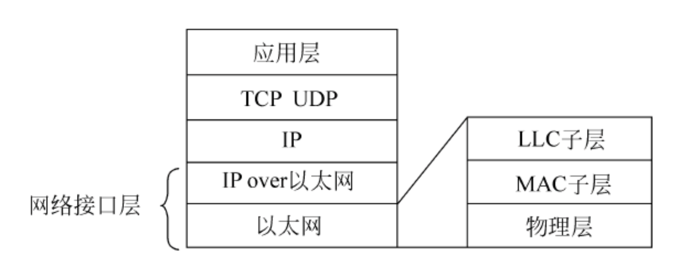


图 1.1　基于局域网的 TCP/IP 体系结构

随着以太网的发展，以太网在局域网市场中已完全取得垄断地位，目前已不存在多种局域网技术并存的问题，而且，LLC 子层是 802 委员会为屏蔽多种局域网之间的差异而提出的，显然不是 DIX Ethernet V2 中的一部分，因此实际基于以太网的 TCP/IP 体系结构删除了 LLC 子层，如图 1.2 所示。

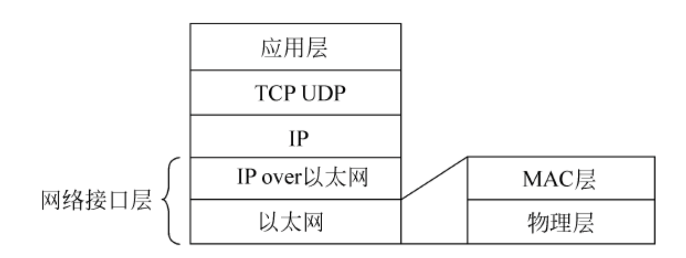


图 1.2　基于以太网的 TCP/IP 体系结构

1.1.3　以太网拓扑结构

在讨论以太网时，不时会提到网络拓扑结构，拓扑（Topology）是拓扑学中研究由点、线组成的几何图形的一种方法，用此方法可以把计算机网络看做是由一组结点和链路组成的几何图形，这些结点和链路所组成的几何图形就是网络的拓扑结构。由于以太网是一个可以由某个单位单独拥有，且允许自主布线的网络，因此，用户对以太网的拓扑结构有较大的选择空间。以太网常见的拓扑结构有总线型、星型、树型和网状型。

1. 总线型拓扑结构

总线型拓扑结构如图 1.3 所示，通常用同轴电缆作为网络中的总线。为了防止反射信号干扰总线上用于传输数据的基带信号，总线两端必须接匹配阻抗。总线型拓扑结构的优点是简单，缺点是连接在总线上的任何一个终端发生故障，都有可能使总线的阻抗发生变化，导致基带信号传输失败。

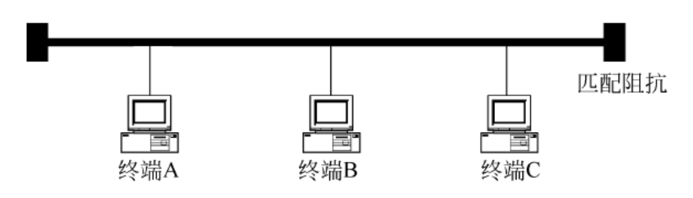


图 1.3　总线型拓扑结构

2. 星型拓扑结构

星型拓扑结构如图 1.4 所示，网络核心设备是物理层的集线器或链路层的交换机，核心设备和终端之间的传输媒体一般为双绞线或光纤，尤其是双绞线作为以太网传输媒体后，其柔软性非常容易满足办公环境下的布线要求，因此出现了一个新的行业——综合布线，并因此将以太网设计和实施分为同等重要的两部分：解决设备之间互连问题的布线系统、实现数据端到端传输及提供应用服务的网络传输系统和应用系统。星型拓扑结构是目前以太网设计中普遍使用的网络结构，当然，实际以太网常常通过级联集线器或交换机将多个星型网络连接在一起。星型拓扑结构的优点是核心设备能够隔离每一个终端，因此，某个终端发生故障或者核心设备用于连接终端的某个端口发生故障，不会影响其他终端之间的通信，这是星型拓扑结构取代总线型拓扑结构的主要原因。

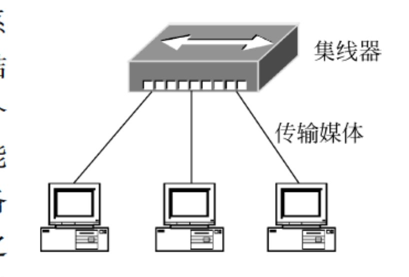
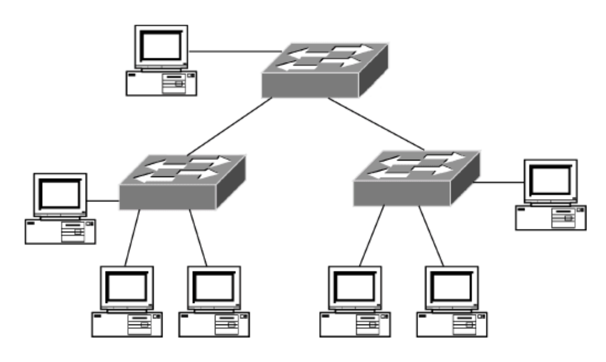


图 1.4　星型拓扑结构

3. 树型拓扑结构

树型拓扑结构如图 1.5 所示，这种拓扑结构实际上就是通过级联交换机或集线器将多个星型拓扑结构连接在一起的网络结构。正常的树型结构要求任何两个终端之间不允许存在环路。


图 1.5　树型拓扑结构

4. 网状型拓扑结构

有容错性要求的以太网为了保证故障情况下终端之间的连通性，往往使交换机之间构成环路，这种在树型拓扑结构上增加环路的拓扑结构称为网状型拓扑结构，如图 1.6 所示。在后面章节中将讨论到，透明网桥的工作原理要求交换机之间不允许存在环路，因此，为了透明网桥能够正常工作，同时又能保证以太网的容错性，必须做到：在网络运行时，通过阻塞某些端口使整个网络没有环路，当某条链路因为故障无法通信时，通过重新开通原来阻塞的一些端口，使网络终端之间依然保持连通性，而又没有形成环路。生成树协议（Spanning Tree Protocol，STP）就是这样一种实现机制。

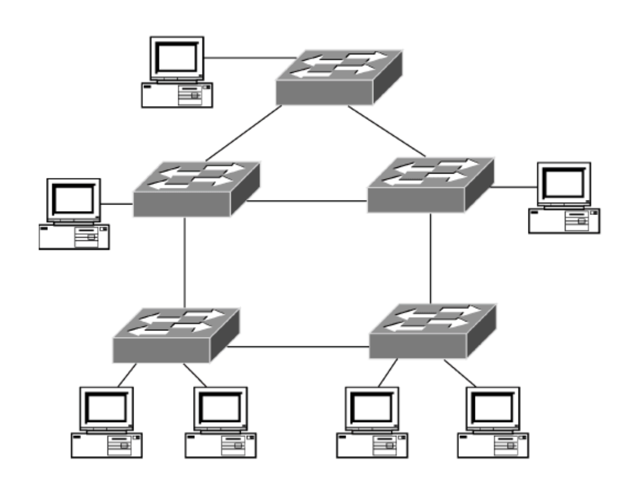


图 1.6　网状型拓扑结构

1.2　以太网从共享到交换

802.3 标准定义的以太网是一种采用总线型拓扑结构，物理层采用曼彻斯特编码，MAC 层用 CSMA/CD 算法解决终端争用总线问题的网络。但随着以太网的发展，无论物理层的编码技术，还是网络拓扑结构都发生了很大变化。本节从总线型以太网开始，讨论以太网从共享式到交换式的发展历程。

1.2.1　总线型以太网

1. MAC 地址

安装在终端中的每一块以太网卡(也称网络适配器)都有唯一的 48 位 MAC 地址，48 位 MAC 地址通常由 6 个用冒号分隔的字节组成，每一个字节用两个十六进制值表示。网卡 MAC 地址在出厂时已经设定，不可更改。48 位 MAC 地址中高 24 位为企业标识符，用于标识以太网卡的生产企业，低 24 位用于区分该企业生产的以太网卡。

MAC 地址分为单播地址、广播地址和组播地址。48 位 MAC 地址中第一个字节的最低位是 I/G 位，该位为 0，表示该 MAC 地址为单播地址，对应单个终端。该位为 1，表示该 MAC 地址为广播或组播地址，对应一组终端。48 位 MAC 地址中第一个字节的次低位是 G/L 位，该位为 0，表示该 MAC 地址是局部地址。该位为 1，表示该 MAC 地址是全局地址，全局地址是指该 MAC 地址全球范围内唯一。

广播地址是 48 位全 1 的地址，用十六进制数表示是 ff:ff:ff:ff:ff:ff(6 个用冒号分隔的全 1 字节)。

组播地址范围是：01:00:5e:00:00:00～01:00:5e:7f:ff:ff。

单播地址是广播和组播地址以外且 I/G 位为 0 的 MAC 地址。

2. MAC 帧结构

按照开放系统互连/参考模型(Open Systems Interconnection/Reference Model,

OSI/RM)定义的网络体系结构,总线型以太网两端接匹配阻抗的单根同轴电缆就是物理链路,在物理链路上传输的是MAC帧的分组,它由数据和用于实现数据传输的地址信息及检测传输过程中出错情况的检错码等组成,总线型以太网的链路层称为MAC层,除了必须实现帧定界、寻址、差错控制等这些基本链路层功能外,还需要实现多点接入网络的接入控制功能。总线只能进行半双工通信,另外,任何一个终端通过总线发送的数据可以被其他所有终端接收,总线的这一特性使其被称为广播信道。

1)帧定界

数据通过总线进行传输时,必须先对数据进行封装,由于总线型以太网的链路层为MAC层,因此,将总线型以太网的链路层封装形式称为MAC帧。由于在总线上通过基带信号传输的是一串二进制比特流,从一串二进制比特流中分离出每一帧,即确定每一帧的起始和结束字节的过程就是帧定界。

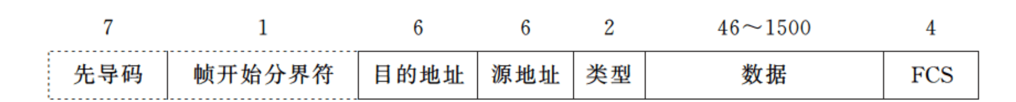


图1.7　MAC帧结构

MAC帧结构如图1.7所示,先导码和帧开始分界符并不是MAC帧的一部分,它们的作用是帮助接收终端完成帧定界的功能。

先导码是由7个二进制比特流模式为10101010的字节组成的一组编码,它的作用是让连接在总线上的终端进行位同步(让接收端能够正确接收MAC帧中的每一位二进制数)。

帧开始分界符为1字节二进制数位流模式为10101011的编码,用于通知接收端该编码后面是MAC帧。这意味着连接在总线上的每一个终端都必须能够通过先导码和帧开始分界符从经过总线传输的一串数字信号中正确定位MAC帧的起始字节(目的地址字段的第一个字节)。

如果连接在总线上的每一个终端都能够正确监测到先导码和帧开始分界符,实现帧定界是没有问题的。由于总线型以太网规定在传输完每一帧后,必须让总线空闲一段时间,而曼彻斯特编码很容易让终端监测到总线从空闲状态转变为发送先导码状态的转换过程,并因此实现帧定界功能。所以,严格地说,总线型以太网的帧定界并不是由MAC层实现,而是由物理层实现的。

2)寻址

由于总线型以太网在同一总线上连接多个终端,通过总线传输的每一MAC帧寻找接收终端的过程就是寻址过程。

MAC帧中的目的地址和源地址字段给出该MAC帧的接收端和发送端的地址,MAC帧携带MAC地址的原因是总线型以太网是一个多点接入网络,允许同时有多个(大于2个)终端或网络互连设备接入总线型以太网,这种情况下,只有携带用于标识接收端目的地址信息的MAC帧,才能确定它的接收终端,并因此实现寻址功能。

当图1.3中终端A发送MAC帧给终端B时,用终端A网卡的MAC地址作为MAC帧的源MAC地址,用终端B网卡的MAC地址作为MAC帧的目的MAC地址,当终端A通过总线发送该MAC帧时,连接在总线上的所有终端都接收到该MAC帧,但都用自己网

卡的 MAC 地址和 MAC 帧中的目的 MAC 地址比较,如果相符,则继续处理,否则将该 MAC 帧丢弃。因此,当一个终端想要给另一个终端发送 MAC 帧时,它必须先获取另一个终端的 MAC 地址,否则,只能以广播方式发送 MAC 帧。

MAC 帧的源 MAC 地址必须是单播地址,目的 MAC 地址可以是单播地址、组播地址或广播地址。对于图 1.3 所示的总线结构,目的 MAC 地址类型对 MAC 帧的发送方式没有影响,无论目的 MAC 地址是单播、组播或广播地址,都同样将 MAC 帧发送到总线上,被连接在总线上的所有其他终端所接收。

目的 MAC 地址类型与终端对接收到的 MAC 帧的处理方式有极大关系,如果是单播 MAC 地址,则只有目的 MAC 地址和其网卡 MAC 地址相符的单个终端接收并处理该 MAC 帧。如果是广播地址,则连接在总线上的所有终端均需接收并处理该 MAC 帧。如果是组播地址,只有属于组播地址所指定的组播组的终端才需要接收并处理该 MAC 帧。组播地址前 25 位是固定不变的,为 01:00:5e:0x(第 4 个字节中只有最高位固定为 0),后 23 位用于标识组播组。

3) 差错控制功能

帧检验序列(Frame Check Sequence,FCS)字段用于接收端检验 MAC 帧在传输过程中是否出错。总线型以太网采用循环冗余检验(Cyclic Redundancy Check,CRC)码对 MAC 帧进行检错,使用的生成多项式如下:

$$G(x) = x^{32} + x^{26} + x^{23} + x^{22} + x^{16} + x^{12} + x^{11} + x^{10} + x^{8} + x^{7} + x^{5} + x^{4} + x^{2} + x + 1$$

32 位帧检验序列(FCS)就是用生成多项式除以由目的地址、源地址、类型、数据和填充字段组合成的二进制比特流后得到的余数。由于 CRC 对检测连续多位二进制数出错非常有效,而且无论用同轴电缆、双绞线还是光纤传输基带信号,有效传输距离内的误码率都很低,因此,总线型以太网在 MAC 层通过帧检验序列(FCS)字段能够检测出绝大多数的传输错误。

接收端一旦通过 FCS 字段检测出 MAC 帧传输过程中发生的错误,则丢弃该 MAC 帧。由于 MAC 层只有检错功能,没有重传机制,因此,一旦发生 MAC 帧传输出错的情况,MAC 层并没有相应的补救措施。

4) MAC 帧其他字段功能

类型字段用于标明数据类型,MAC 帧所封装的数据可以是 IP 分组,也可以是 ARP 请求及其他类型的数据,包含不同类型数据的 MAC 帧需要提交给不同的进程进行处理,类型字段就用于接收端选择和数据的类型相对应的进程。

数据字段用于传输数据,和其他字段不同,数据字段是真正承载高层协议要求传输的数据,而其他字段只是用于保证数据的正确传输,因此,把数据字段称做 MAC 帧的净荷字段(也称载荷字段),数据字段的长度是可变的。MAC 帧的这种组织结构很容易让终端从 MAC 帧中分离出数据。

MAC 帧有着严格的长度限制,它的长度必须在 64 字节和 1518 字节之间,由于其他字段占用了 18 个字节(6 个字节源 MAC 地址+6 个字节目的 MAC 地址+2 个字节类型+4 个字节帧检验序列),数据字段长度应该在 46 字节和 1500 字节之间,但高层协议要求传输的数据的长度应该是任意的,一旦数据的字节数不足 46 字节,就需要用填充字段将 MAC 帧的长度延长到 64 字节,由此可以推出填充字段的长度在 0 字节和 46 字节之间。

对 MAC 帧规定长度上限可以理解,因为一是每个接收终端的缓冲器空间有限,二是不能允许某个终端通过发送无限长的 MAC 帧独占总线,三是一旦长度很长的 MAC 帧传输出错,重新传输的代价太大。但对 MAC 帧规定长度下限却不好理解,下面章节将详细讨论对 MAC 帧规定长度下限的意义。

3. CSMA/CD 操作过程

1) CSMA/CD 工作原理

如果图 1.3 中某一个终端想要发送数据,通过 CSMA/CD 算法解决和其他终端争用总线的问题。CSMA/CD 的中文是载波侦听(Carrier Sense,CS)、多点接入(Multiple Access,MA)/冲突检测(Collision Detection,CD),它用于解决多个终端争用总线的机制如下。

(1) 先听再讲:想发送数据的终端必须确定总线上没有其他终端正在发送数据后,才能开始往总线上发送数据,即先要侦听总线上是否有载波(这里的载波指其他终端发送 MAC 帧时产生的高低电平有规律跳变的电信号,不是调制过程中使用的特定频率的正弦信号),在确定总线空闲(无载波出现)的情况下,才能开始发送数据。一旦开始发送数据,随着电信号在总线上传播,总线上所有其他终端都能侦听到载波存在,这就是先听(侦听总线载波)再讲(发送数据)。

(2) 等待帧间最小间隔:并不是一侦听到总线空闲就立即发送数据,而必须侦听到总线空闲一段时间(称为帧间最小间隔:IFG,10Mb/s 以太网的帧间最小间隔为 9.6μs)后,才能开始发送数据。这样做的目的有三:一是如果接连两 MAC 帧的接收终端相同,必须在两帧之间给接收终端一点用于腾出缓冲器空间的时间。二是一个想连续发送数据的终端,在发送完当前帧后,不允许接着发送下一帧,必须和其他终端公平争用发送下一帧的机会。三是总线在发送完一 MAC 帧后,必须回到空闲状态,以便在发送下一 MAC 帧时,能够让连接在总线上的终端正确监测到先导码和帧开始分界符。

(3) 边讲边听:一旦某个终端开始发送数据,其他终端都能侦听到载波,即使这些终端中存在想发送数据的终端,它也必须等待,直到总线空闲一段时间(由 IFG 确定)后,才能开始发送数据。但可能存在这样一种情况:两个终端都想发送数据,因此都开始侦听总线,当当前帧发送完毕时,两个终端同时侦听到总线空闲,并在总线空闲状态持续一段时间后,同时发送数据,这样,两个终端发送的电信号就会叠加在总线上,导致冲突发生。其实,由于电信号经过总线传播需要时间,如果两个终端相隔较远,即使一个终端开始发送数据,在电信号传播到另一个终端前,另一个终端仍然认为总线空闲,因此,即使不是同时开始侦听总线,只要两个终端开始侦听总线的时间差在电信号传播时延内,仍然可能发生冲突。因此,某个终端开始发送数据后,必须一直检测总线上是否发生冲突,如果检测到冲突发生,停止正常的数据传输,发送 4 字节或 6 字节长度的阻塞信号(也称干扰信号),加重冲突情况,使所有发送数据的终端都能检测到冲突情况的发生。这就是边讲(发送数据)边听(检测冲突是否发生)。检测冲突是否发生的方法很多,其中比较简单的一种是边发送边接收,并将接收到的数据和发送的数据进行比较,一旦发现不相符的情况,表明冲突发生。

(4) 退后再讲:一旦检测到冲突发生,停止数据发送过程,延迟一段时间后,再开始侦听总线。两个终端的延迟时间必须不同,否则可能进入发送→冲突→延迟→侦听→发送→冲突这样的循环中。如果两个终端的延迟时间不同,延迟时间短的终端先开始侦听总线,在

侦听到总线空闲并持续空闲一段时间后，开始发送数据，当延迟时间长的终端开始侦听总线时，另一个终端已经开始发送数据，它必须等待总线空闲后，才可以开始发送过程，CSMA/CD 操作过程如图 1.8 所示。

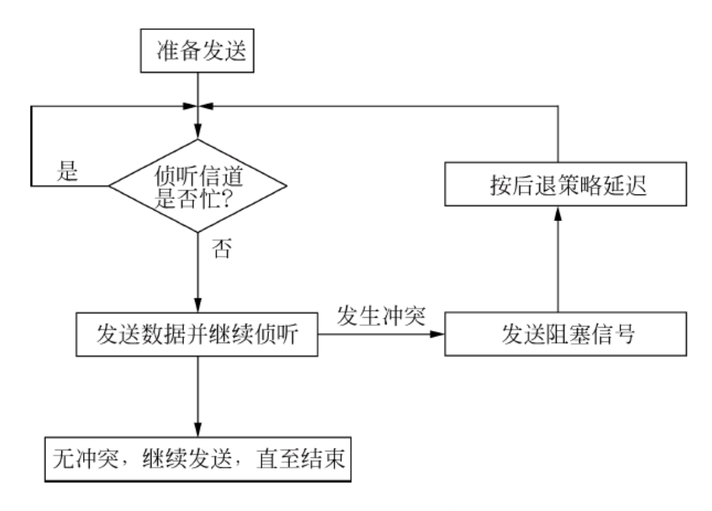


图 1.8 CSMA/CD 操作过程

2）后退算法

发生冲突后，两个终端的延迟时间必须不同，否则将再次发生冲突。如何使两个终端产生不同的延迟时间呢？以太网采用称为截断二进制指数类型的后退算法，算法的基本思路如下：

① 确定参数 K。开始时 $K=0$，每发生一次冲突，K 就加 1，但 K 不能超过 10，因此，$K=\mathrm{MIN}[$冲突次数$,10]$。

② 从整数集合$[0,1,\cdots,2^K-1]$中随机选择某个整数 r。

③ 根据 r，计算出后退时间 $T=r\times t$（t 是最大往返时延，对于 10Mb/s 以太网，$t=51.2\mu s$）。

④ 如果连续重传了 16 次都检测到冲突发生，则终止传输，并向高层协议报告。

一旦两个终端发生冲突，每一个终端单独执行后退算法，在计算延迟时间时，对于第一次冲突，$K=1$，两个终端各自在[0,1]中随机挑选一个整数，由于只有 2 种挑选结果，两个终端挑选相同整数的概率为 50%。如果两个终端在第一次发生冲突后挑选了相同整数，则将再一次发生冲突。当检测到第二次冲突发生时，两个终端各自在[0,1,2,3]中随机挑选整数，由于选择余地增大，两个终端挑选到相同整数的概率降为 25%。随着冲突次数不断增加，两个终端产生相同延迟时间的概率不断降低。当两个终端的延迟时间不同时，选择较小延迟时间的终端先成功发送数据。

截断二进制指数类型的后退算法是一种自适应后退算法，为了提高总线的利用率，在发生冲突时，最好做到：①同时参与争用总线行动的终端的延迟时间不能相同；②最小的且与其他终端的延迟时间不同的延迟时间最好为 0。当参与争用总线行动的终端少时（最少为 2 个），有 50%的可能是一个终端选择 0 延迟时间，另一个终端选择 51.2μs 的延迟时间，选择 0 延迟时间的终端可以立即侦听总线，并在总线空闲时发送数据，这对提高总线利用率当然有益。但当有多个终端（假定为 100 台）参与争用总线的行动时，在第一次冲突发生时，其中一个终端选择 0 延迟时间，其余 99 个终端选择 51.2μs 延迟时间的概率实在太小。但

随着冲突次数的不断增多，整数集合的不断扩大，很有可能在发生16次冲突前，有一个终端选择了整数r，它和所有其他终端选择的整数不同，且小于所有其他终端选择的整数。

3）捕获效应

截断二进制指数类型的后退算法在两个终端都想连续发送数据的情况下，有可能导致一个终端长时间内一直争到总线发送数据，而另一个终端长时间内一直争不到总线发送数据，即所谓的捕获效应。

如图1.9所示，当两个终端同时想长时间发送数据时，都去侦听总线，当总线空闲后（总线持续空闲IFG规定的时间），两个终端同时向总线发送数据，导致冲突发生，用截断二进制指数类型算法求出后退时间，假定终端A选择的后退时间为$0\times t$，而终端B选择的后退时间为$1\times t$，终端A的第1帧数据发送成功。由于终端A有大量数据需要发送，在发送完第1帧数据后，紧接着发送第2帧数据，但必须通过争用总线过程才能获得发送第2帧数据的机会。当终端A和终端B又侦听到总线空闲，并又同时发送数据，导致冲突再次发生时，对于终端A而言，由于它是因为发送第2帧数据而导致发生的第1次冲突，因此冲突次数$K=1$，在整数集合[0,1]之间随机挑选一个整数r，而对于终端B而言，它是因为发送第1帧数据而导致发生的第2次冲突，冲突次数$K=2$，在整数集合[0,1,2,3]中随机挑选一个整数r'，显然，$r<r'$的概率更大，使得终端A又一次成功发送第2帧数据。根据终端B选择的延迟时间大小和终端A的MAC帧长度，有可能在终端B的后退时间内，终端A已成功发送若干MAC帧。但当终端B再次开始侦听总线并试图发送数据时，又将和终端A发生冲突，对于终端A，仍然在整数集合[0,1]中随机挑选一个整数r，而终端B将在整数集合[0,1,2,3,4,5,6,7]中随机挑选一个整数r'，$r<r'$的概率比前一次更大，又导致终端A发送成功。这样，就使得终端A长时间通过总线发送数据，而终端B一直得不到发送数据的机会。

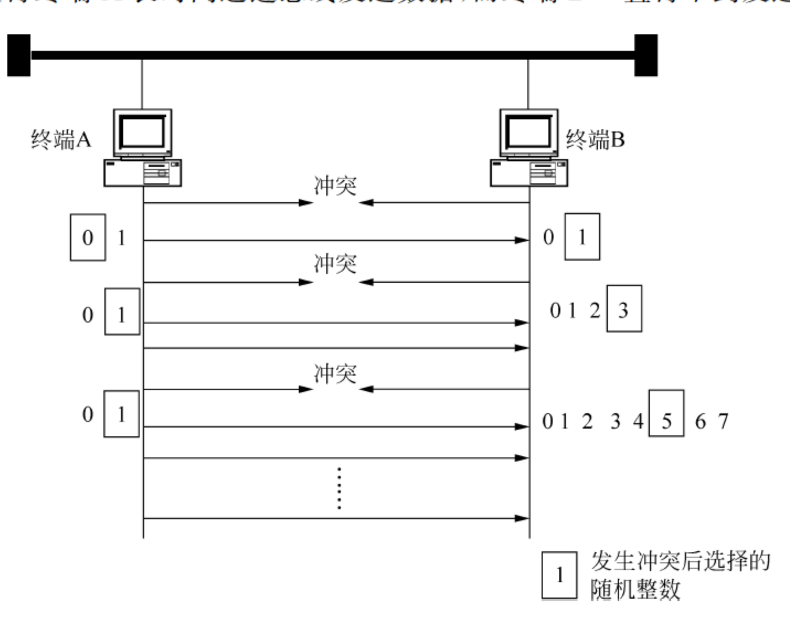


图1.9　捕获效应示意图

捕获效应是非常严重的问题，如果不是以太网从共享发展成交换，端口通信方式从半双工发展为全双工，捕获效应将成为厂家需要面对的一个重大问题。后面章节将讲到，随着交

换式以太网的兴起和以太网交换机端口与主机之间广泛采用全双工通信方式，捕获效应问题自然消失。

4. 冲突域直径和最短帧长

1）冲突域直径

在前面讨论 MAC 帧字段时已经讲到，接收端对接收到的 MAC 帧进行差错检验，如果发现接收到的 MAC 帧在传输过程中出错，接收端将丢弃该 MAC 帧，否则，将 MAC 帧提交给高层协议。对于接收端而言，无论接收到的 MAC 帧正确与否，都不会向发送端提供任何有关该 MAC 帧是否成功接收的信息，因此，发送端只要将 MAC 帧发送成功，就认为对该 MAC 帧的处理已经完成，即使该 MAC 帧在传输过程中出错，由高层协议（如 TCP），而不是由 MAC 层进行差错控制。这就要求经过以太网传输的 MAC 帧的出错率必须保持在很低的水平，避免在终端之间重复传输传输层报文情况的发生。对于总线型以太网，最有可能导致 MAC 帧传输出错的原因是发生冲突，因此，发送端一旦检测到冲突发生，就不能认为发送成功，而是后退一段时间后，再次重发。那么如何确保发送端能够检测到任何情况下发生的冲突呢？

在总线型以太网中，只允许一个终端发送数据，一旦有两个或两个以上终端同时发送数据，就会发生冲突，因此，将具有这种传输特性的网络所覆盖的地理范围，称为冲突域，将同一冲突域中相距最远的两个终端之间的物理距离称为冲突域直径。在下面的讨论中，不是用距离而是用时间来标识冲突域直径，是因为在知道电信号传播速度的情况下，传播时间和传播距离是可以相互换算的。假定同轴电缆的长度为 L，电信号传播速度为 V，则传播时间 $T=L/V$。如果在真空中，电信号传播速度等于光速 c。在电缆中由于阻抗的因素，电信号的传播速度约为(2/3)c，因此，$T=3L/2c$。反过来，确定了传播时间 T，也可得出电缆长度 $L=(2/3)c\times T$。

2）中继器扩展电信号传播距离

电信号通过电缆传播会产生衰减，衰减程度与电缆的长度成正比，因此，单段电缆不允许很长，表 1.1 中给出了不同传输媒体类型单段电缆的长度。为了扩大冲突域直径，必须使用电缆连接设备——中继器（或转发器）。中继器是一个物理层设备，它的功能是将衰减后的电信号再生，即放大和同步，图 1.10 给出中继器再生基带信号的过程。中继器将一端接收到的已经衰减的电信号经放大、同步后从另一端输出的过程需要时间，因此，在使用中继器互连电缆的冲突域中，不能简单地根据作为冲突域直径的时间 T，推算出物理距离 $L=(2/3)c\times T$，而必须考虑中继器再生电信号需要花费的时间。如果每一个中继器的延迟时间为 T'，冲突域中有 N 个中继器，根据作为冲突域直径的时间 T，可大致推算出冲突域直径的物理距离 $L=(2/3)c\times(T-N\times T')$。

中继器是物理层互连设备，理论上可以通过中继器的信号再生功能将冲突域无限扩大，即经过中继器互连的同轴电缆总长不受限制。

3）MAC 帧的最短帧长

为了保证发送端能够检测到任何情况下发生的冲突，MAC 帧的发送时间和冲突域直径之间存在相互制约。而 MAC 帧的长度和总线的传输速率又决定了 MAC 帧的发送时间，因此，冲突域直径和 MAC 帧的最短帧长之间存在相互制约，下面通过图 1.11 讨论一下它们之间的关系。

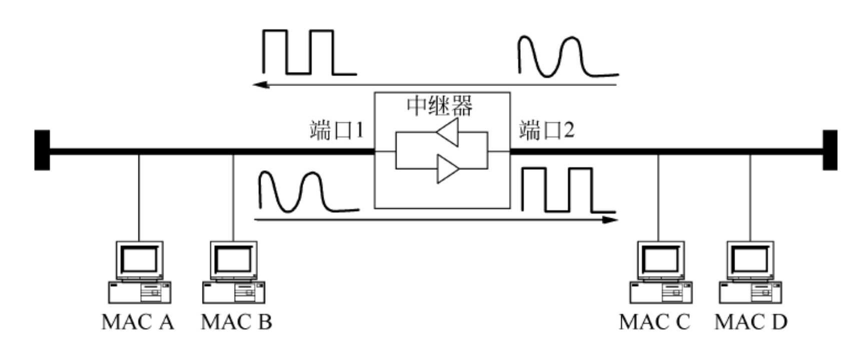


图 1.10 中继器再生基带信号的过程

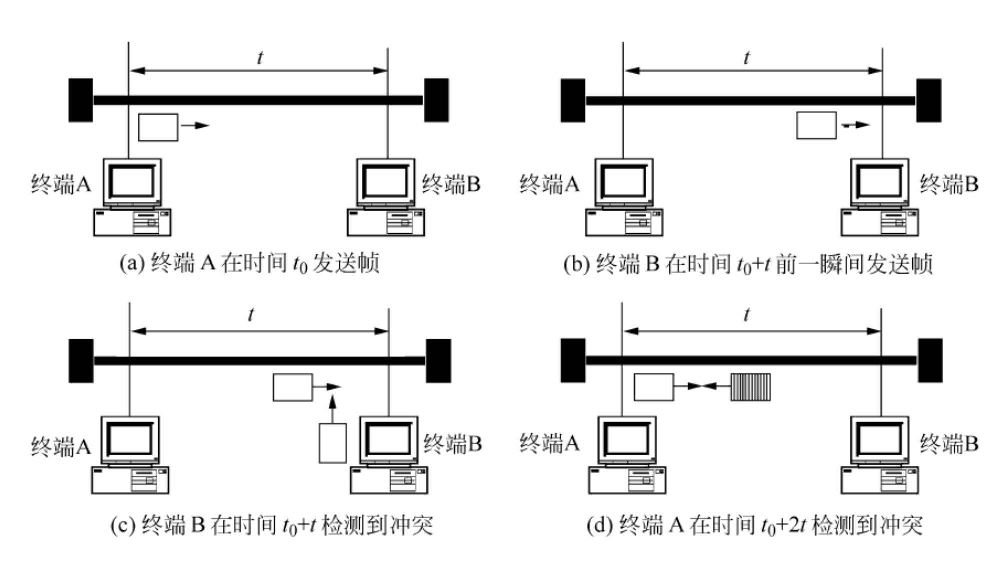


图 1.11 冲突域直径和最短帧长之间关系

图 1.11 中假定已经知道了冲突域直径是时间 t，这意味着电信号从终端 A 传播到终端 B 所需要的时间为 t（电信号传播过程中可能经过若干中继器），那么如何确定最短帧长呢？

假定终端 A 在时间 t_0 开始发送 MAC 帧，在 t_0+t 前一瞬间，终端 B 由于侦听到总线空闲也开始发送数据（见图 1.11 (b)）。当然，终端 B 立即检测到冲突发生（见图 1.11 (c)）。它一方面停止发送 MAC 帧，另一方面通过发送阻塞信号来强化冲突，方便终端 A 对冲突的检测。但终端 B 发送的电信号必须经过时间 t 才能到达终端 A（见图 1.11 (d)），和终端 A 发送的电信号叠加，使终端 A 能够检测到冲突发生。由于终端 A 发送每一个 MAC 帧的过程是边发送边检测冲突是否发生，因此，为了确保能够检测到任何情况下发生的冲突，终端 A 发送 MAC 帧的时间不能小于 $2t$，因此，将发送时间为 $2t$ 的 MAC 帧长度称为最短帧长，如果最短帧长为 M，网络传输速率为 S，则 $M/S=2t$，求出 $M=2t\times S$。10Mb/s 以太网标准规定 $t=25.6\mu s$，$2t=51.2\mu s$，$S=10Mb/s$，求出 MAC 帧最短帧长$=51.2\times10^{-6}\times10\times10^{6}=512b=64B$。64B 最短帧长的含义是：在确定冲突域直径为 $25.6\mu s$ 的前提下，发送端只有保证每一帧的发送时间$\geq51.2\mu s$，才能检测到任何情况下发生的冲突。$2t$ 称为争用期，也称为冲突窗口，任何一个终端只有在冲突窗口内没有检测到冲突发生，才能保证该次发送不会发生冲突。

4）最短帧长对高速以太网冲突域直径的限制

如果没有中继设备，冲突域两端之间直接用电缆连接，$25.6\mu s$ 的冲突域直径转换成的

物理距离$=25.6\times10^{-6}\times2\times10^{8}$($2c/3=2\times10^{8}$m/s)$=5120$m,但无论粗同轴电缆,还是细同轴电缆,单段电缆的长度都不可能达到5120m,如表1.1所示,因此,必须使用中继器,使用中继器后的冲突域直径的物理距离和冲突域两端之间通路中的中继器数量及中继器实现信号再生所需要的时间有关。表1.1给出了不同传输媒体在25.6μs传播时间能够达到的物理距离,即转换成物理距离的冲突域直径。需要强调的是:表1.1是标准推荐的冲突域直径的物理距离,它不仅需要考虑中继器信号再生过程所需要的时间,还须有一定的冗余,因此,小于极端条件下计算出的物理距离。

表1.1　各种类型电缆的物理距离

传输媒体类型	中继器数量	单段电缆长度/m	冲突域直径/m
粗同轴电缆	4	500	2500
细同轴电缆	4	185	925
双绞线	4	100	500

在明白了最短帧长和冲突域直径之间的关系后,就会发现以太网发展过程中遇到的诸多困难。电信号传播时间与终端发送数据的速率无关,基本上只和传输距离和中间经过的中继器数量有关,当终端网卡的传输速率从10Mb/s上升到100Mb/s时,如果保持冲突域直径不变(仍然为25.6μs),发送端发送MAC帧的时间必须大于25.6μs×2=51.2μs,计算出最短帧长$=(51.2\times10^{-6})\times(100\times10^{6})=5120$b,即640B。如果为了兼容,要求最短帧长不变,仍为64B,则冲突域直径必须缩小到以100Mb/s传输速率发送512位二进制数所需时间的一半($t=M/(2S)$),即$(512/(100\times10^{6}\times2))$秒$=2.56$μs,将其转换成物理距离的话,大约在200m左右(考虑中间存在中继器的情况)。100Mb/s以太网选择的最短帧长和10Mb/s以太网兼容,但转换成物理距离的冲突域直径却降低到216m,图1.12是标准推荐的100Mb/s以太网的连接方式。216m是冲突域两端之间通路存在2个中继器的情况下,电信号在2.56μs时间内所能传播的最大物理距离。

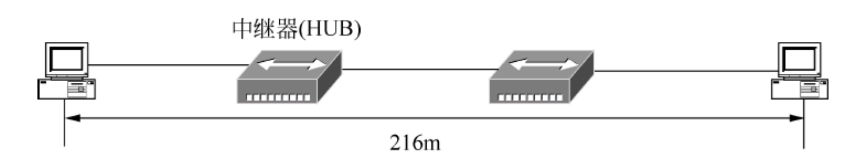


图1.12　100Mb/s以太网的连接模式

当以太网传输速率从100Mb/s发展到1000Mb/s时,如果维持最短帧长不变,仍为64B,则冲突域直径将缩小到以1000Mb/s传输速率发送512位二进制数所需时间的一半,即$(512/(1000\times10^{6}\times2))$秒$=0.256$μs。最短帧长和冲突域直径之间矛盾更加突出。这种情况下,转换成物理距离的冲突域直径将下降为50m左右,网络将失去实际意义。因此,1000Mb/s以太网将最短帧长选择为640B,这样,冲突域直径可以提高到以1000Mb/s传输速率发送5120位二进制数所需时间的一半,即$(5120/(1000\times10^{6}\times2))$秒$=2.56$μs,仍然能够将转换成物理距离的冲突域直径维持在200m左右。

1000Mb/s以太网扩大最短帧长的方法有两种:一是将多个帧长小于640B的MAC帧集中起来当做一个MAC帧发送,保证发送时间大于2.56μs。另一种方法是如果发送的MAC

帧的长度小于640B，网卡在发送完MAC帧后，继续发送填充数据，保证每次发送时间大于2.56μs，通过这两种方法既保证了将最短帧长扩大到640B，又和10Mb/s、100Mb/s以太网兼容。

5. 集线器和星型以太网结构

自从出现把双绞线作为传输媒体的以太网标准，人们开始广泛采用集线器（Hub）来互连终端。集线器是一个多端口中继器，端口支持的传输媒体类型通常为双绞线，因此，用集线器连接终端方式构建的以太网仍然是一个共享式以太网，即整个以太网是一个冲突域，图1.13是用集线器互连终端的网络结构和集线器工作原理图。

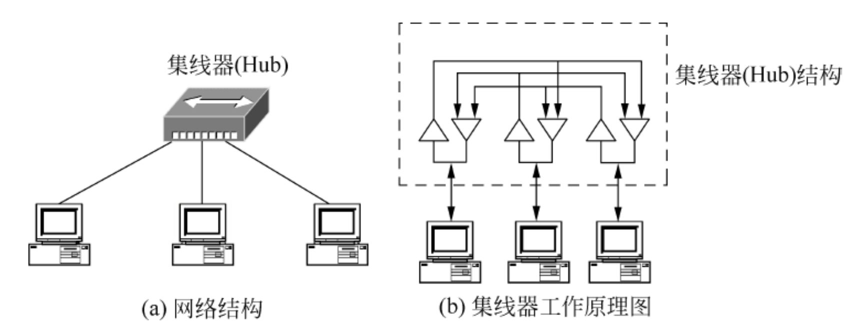


图1.13　集线器互连终端的网络结构和集线器工作原理图

从图1.13中可以看出：虽然连接终端的双绞线电缆分别有一对双绞线用于发送，一对双绞线用于接收，但一旦某个终端发送数据，发送的数据将传播到所有终端的接收线上，因此，任何时候仍然只允许一个终端发送数据。可以说集线器只改变了以太网的拓扑结构，将以太网从总线型变为星型，但终端通过争用总线传输数据的实质没有改变。可以将集线器想象成缩成一个点的总线，把这种物理上的星型网络当做逻辑上的总线型网络，即从物理连接方式看是星型，但从信号传输方式看，仍然和总线型以太网相同。因此，仍然可以将由集线器连接终端构成的冲突域作为单个广播信道。

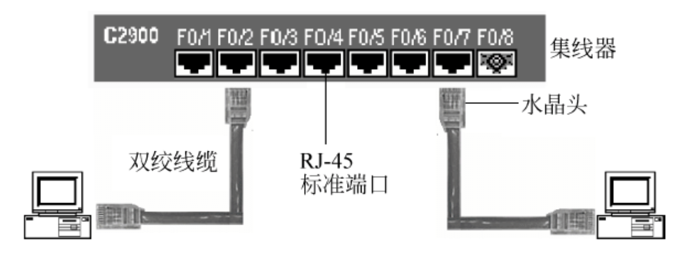


图1.14　集线器互连终端示意图

集线器互连终端的方式如图1.14所示，将两端连接水晶头的双绞线缆的一端插入集线器RJ-45标准端口。另一端插入PC网卡RJ-45标准端口，RJ-45是双绞线接口标准。多台集线器可以通过双绞线缆串接在一起，根据以太网标准，10Mb/s的集线器最多串接4个，冲突域直径为500m。100Mb/s的集线器最多串接2个，冲突域直径为216m，图1.12就是串接两个集线器的网络结构。

从前面讨论中可以得出：CSMA/CD算法虽然解决了冲突域内终端争用总线的问题，但严格限制了同一冲突域内的终端数量和两个终端之间的最大物理距离，这使得以太网的

应用前景蒙上了阴影，尤其在异步传输模式(Asynchronous Transfer Mode，ATM)技术出现后，人们纷纷预测 ATM 网络将在不久之后取代以太网。

6. 例题解析

【例 1.1】 根据 CSMA/CD 工作原理，下述情况中需要提高最短帧长的是________。

A. 网络传输速率不变，冲突域最大距离变短

B. 冲突域最大距离不变，网络传输速率变高

C. 上层协议使用 TCP 概率增加

D. 在冲突域最大距离不变的情况下，减少线路中的中继器数量

【解析】 最短帧长$=2\times T\times S$，T 是化作时间的冲突域直径，即电信号经过冲突域最大距离传播所需要的时间，S 是网络传输速率。

A 情况中，T 减小，S 不变，最短帧长应该减小，不是提高。

B 情况中，T 不变，S 变高，最短帧长应该提高。

C 情况中，底层网络的工作过程应该和高层协议无关，这也是分层的主要原因，因此，上层使用 TCP 的概率和最短帧长无关。

D 情况中，由于线路中的中继器数量减少，在冲突域最大距离不变的情况下，电信号经过冲突域最大距离传播所需要的时间 T 减少，最短帧长应该减小。

综合以上分析，正确答案是 B。

【例 1.2】 在一个采用 CSMA/CD 的网络中，传输介质是一根完整的电缆，传输速率为 1Gb/s，电缆中的信号传播速度为(2/3)c($c=3\times10^8$m/s)，若最小帧长减少 800b，则相距最远的两个站点之间的距离至少需要________。

A. 增加 160m　　B. 增加 80m　　C. 减少 160m　　D. 减少 80m

【解析】 根据公式：最短帧长$=2\times T\times S$，求出作为冲突域直径的时间差 $\Delta T=$最小帧长差$/(2\times S)=800/(2\times10^9)=4\times10^{-7}$s，冲突域直径的距离差=冲突域直径的时间差 $\Delta T\times$电信号传播速度$=4\times10^{-7}\text{s}\times3\times10^8\text{m/s}=80\text{m}$。正确答案是减少 80m，选 D。

【例 1.3】 某局域网采用 CSMA/CD 协议实现介质访问控制，数据传输速率为 10Mb/s，主机甲和主机乙相距 2km，信号传播速度为 200000km/s，请回答下列问题并给出计算过程。

(1) 若主机甲和主机乙发送数据时发生冲突，从开始发送数据到两个主机均检测到冲突发生的最短和最长时间分别是多少？(假定主机甲和主机乙发送数据期间，其他主机不发送数据)

(2) 若网络不存在任何冲突和差错，主机甲以以太网标准允许的最长数据帧(1518B)向主机乙发送数据，一旦主机乙成功接收当前数据帧，主机甲立即发送下一帧，问主机甲的有效数据传输速率是多少？(不考虑 MAC 帧的先导码)

【解析】 (1) 主机甲和主机乙同时发送数据的情况下，所需时间最短，为端到端传播时延，等于 $2/200000=1\times10^{-5}$s。一方发送的数据到达另一方时，另一方才开始发送数据的情况下，所需时间最长，为端到端传播时延的两倍，等于 $2\times(2/200000)=2\times10^{-5}$s。

(2) 主机甲两数据帧的发送间隔$=(1518\times8)/(10\times10^6)+(2/200000)=1.2244\times10^{-3}$，有效数据传输速率等于间隔时间内实际发送的字节数/间隔时间$=(1518\times8)/$

(1.2244×10^{-3})=9.92Mb/s。

1.2.2　透明网桥与冲突域分割

1. 网桥分割冲突域原理

如果想要扩大以太网的作用范围，必须能够将一个大型以太网分割成若干个冲突域，并用一种设备将多个冲突域互连在一起，这种互连多个冲突域的设备就是网桥。

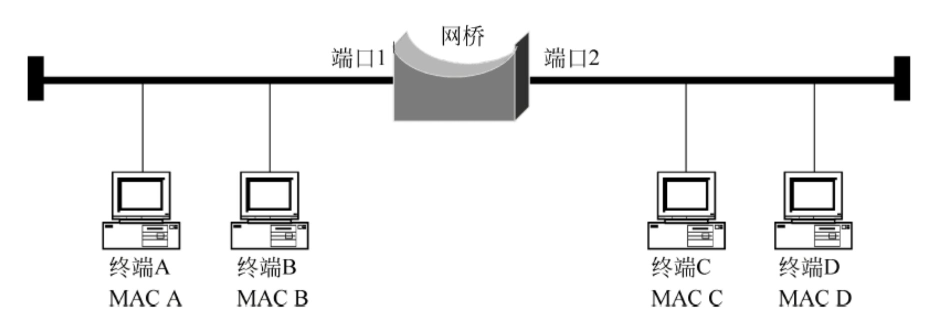


图 1.15　用双端口网桥互连两个冲突域的以太网结构

图 1.15 是一种用双端口网桥将两个冲突域互连成一个以太网的结构，终端 A、终端 B 和网桥端口 1 构成一个冲突域，终端 C、终端 D 和网桥端口 2 构成一个冲突域。和中继器不同，网桥不会将从一个端口接收到的电信号经放大、整形后从另一个端口发送出去。网桥实现电信号隔断和实现位于不同冲突域的终端之间通信功能的原理如图 1.16 所示。

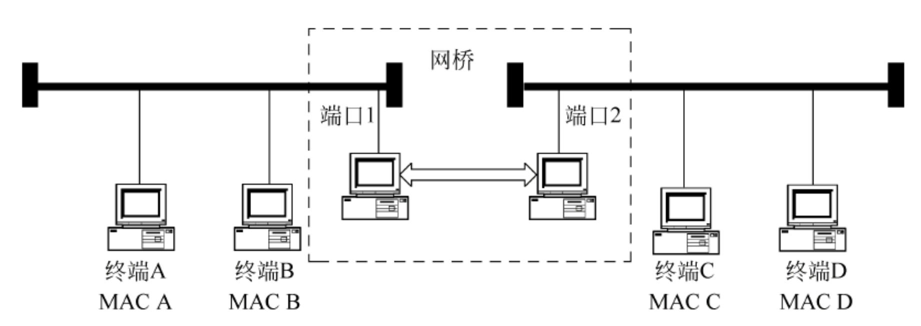


图 1.16　网桥实现电信号隔断并在不同冲突域之间转发 MAC 帧的原理

一个网络成为单个冲突域是因为连接在网络中的任何一个终端所发送的电信号都被传播到整个网络中，中继器虽然从物理上将电缆分割成了两段，但连接在其中一段电缆上的终端所发送的电信号通过中继器传播到另一段，只是中继器在将电信号从一个端口连接的电缆传播到另一个端口连接的电缆时，还将已经衰减的电信号放大、整形、同步，还原成标准的基带信号。虽然网桥和中继器的物理连接方式一样，但网桥完全隔断了电信号的传播通路，对于图 1.15 所示的连接方式，电信号只能在单段电缆上传播，一段电缆上的电信号无法通过网桥传播到另一段电缆上。从电信号传播的角度看，通过网桥连接的两段电缆完全是相互独立的两个冲突域，双端口网桥将网络分割为两个相互独立的总线型以太网，如图 1.16 所示。

在同一个冲突域中，网桥端口和其他终端的功能是一样的，一方面，它也接收其他终端经过总线发送的 MAC 帧。另一方面，它也通过连接的总线发送数据，发送数据时，同样需

要遵循 CSMA/CD 算法，在侦听到总线空闲并维持 IFG 所规定的时间间隔后，才能开始数据发送。为了实现位于不同冲突域的两个终端之间的通信功能，网桥能够从一个端口接收 MAC 帧，再从另一个端口将 MAC 帧转发出去。但切记，当网桥从另一个端口转发 MAC 帧前，它必须先侦听另一个端口所连总线是否空闲，只有在另一个端口所连总线空闲的情况下，网桥才开始转发 MAC 帧。这表明：①网桥必须有缓冲器空间来存储因另一个端口所连总线忙而无法及时转发的 MAC 帧；②两个冲突域可以同时进行数据传输而不会发生冲突，比如图 1.15 中的终端 A 和终端 B、终端 C 和终端 D 就允许同时进行数据传输。同一冲突域中，由于 N 个终端共享总线带宽 M，在不考虑因为冲突导致的带宽浪费的情况下，每一个终端平均分配 M/N 带宽。由于网桥每一个端口连接的冲突域都是独立的，因此，对于 N 个端口的网桥，当每一个端口连接的冲突域的带宽为 M 时，总的带宽是 $N\times M$。

2. 网桥根据转发表转发 MAC 帧

如果是位于同一冲突域的两个终端之间进行数据传输，如图 1.15 中的终端 A 向终端 B 发送数据，网桥连接该冲突域的端口虽然也接收到该 MAC 帧，但丢弃该 MAC 帧，不对该 MAC 帧作任何处理。如果是位于某个冲突域的终端向位于另一个冲突域的终端发送数据，如图 1.15 中的终端 A 向终端 C 发送数据，网桥从连接发送终端所在冲突域的端口接收 MAC 帧，在另一个端口所连的总线空闲的情况下，通过另一个端口将该 MAC 帧转发出去。问题在于网桥如何得知哪些 MAC 帧是需要转发的，哪些 MAC 帧是可以丢弃的呢。

这个功能通过网桥中的转发表（也称 MAC 地址表）来实现，表 1.2 是图 1.15 中网桥所建立的转发表，转发表中的每一项称为转发项，转发项给出的 MAC 地址是某个终端所安装的网卡上的 MAC 地址，对应的转发端口表明该 MAC 地址所对应的终端连接在转发端口所连的冲突域上。例如转发表中其中一项转发项的 MAC 地址是 MAC A，转发端口为端口 1，这表明和 MAC A 这个 MAC 地址关联的终端（终端 A）连接在端口 1 所连的冲突域上。

表 1.2 转发表

MAC 地址	转发端口
MAC A	端口 1
MAC B	端口 1
MAC C	端口 2
MAC D	端口 2

网桥中有了如表 1.2 所示的转发表后，就能够轻易解决确定从一个端口接收到的 MAC 帧是否需要从另一个端口转发出去的问题。由于每一个 MAC 帧都携带源 MAC 地址和目的 MAC 地址，源 MAC 地址给出发送终端的 MAC 地址，而目的 MAC 地址给出接收终端的 MAC 地址，当网桥从一个端口接收到 MAC 帧，它根据 MAC 帧携带的目的 MAC 地址去查找转发表，假定在转发表中找到一项转发项，该转发项的 MAC 地址和 MAC 帧的目的 MAC 地址相同，而该转发项的转发端口为 X，如果端口 X 就是接收到该 MAC 帧的端口，意味着发送该 MAC 帧的终端和接收该 MAC 帧的终端位于同一个冲突域，即网桥端口 X 所连的冲突域，网桥无须对该 MAC 帧作任何处理。如果端口 X 不是网桥接收到该 MAC 帧的端口，意味着接收终端连接在端口 X 所连接的冲突域上，而且和发送终端不在同一个冲突域，网桥必须通过

CSMA/CD 算法成功地通过端口 X 转发该 MAC 帧。如果某项转发项的 MAC 地址和 MAC 帧的目的 MAC 地址相同,表示该转发项和该 MAC 帧的目的 MAC 地址匹配。

3. 网桥工作流程

图 1.17 给出了网桥地址学习和 MAC 帧转发过程。当网桥从端口 X 接收到一个 MAC 帧,意味着端口 X 和该 MAC 帧的发送终端位于同一个冲突域,网桥就可以在转发表中添加一项,该项的 MAC 地址为该 MAC 帧携带的源 MAC 地址,而转发端口为网桥接收该 MAC 帧的端口 X。因此,地址学习过程如下,一旦网桥接收到 MAC 帧,就用 MAC 帧的源 MAC 地址匹配转发表,一旦在转发表中找到匹配的转发项,用接收该 MAC 帧的端口取代转发项中的转发端口,刷新定时器。如果在转发表中找不到匹配的转发项,增加一项转发项,MAC 帧源 MAC 地址为该转发项的 MAC 地址,接收该 MAC 帧的端口为转发项中的转发端口,启动定时器。

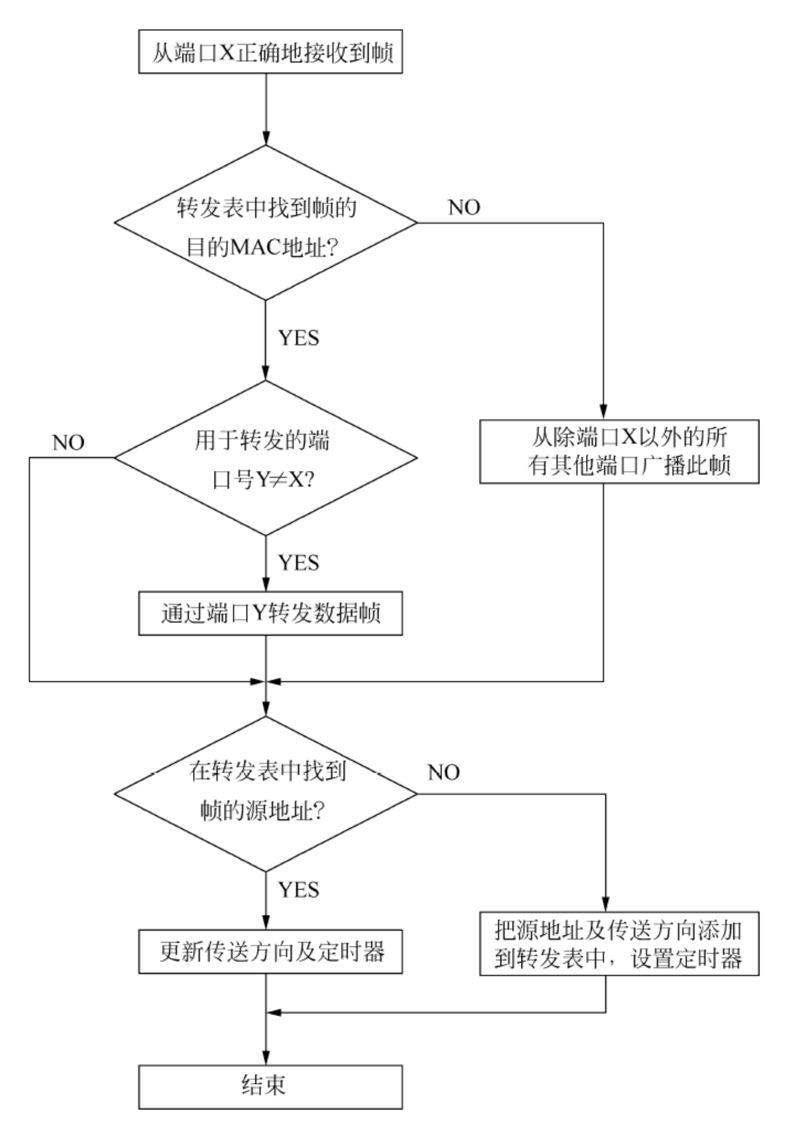


图 1.17　网桥地址学习和 MAC 帧转发过程

只有当网桥所连的两个冲突域上的所有终端均发送了 MAC 帧，网桥才能完整建立如表 1.2 所示的转发表。如果网桥刚初始化，对接收到的第一个 MAC 帧作何处理？或者虽然转发表中已有若干项，但就是没有接收到的 MAC 帧所携带的目的 MAC 地址所匹配的项，那又将如何？

为实现连接在不同冲突域的终端之间的 MAC 帧传输，网桥接收到 MAC 帧后，用 MAC 帧的目的 MAC 地址匹配转发表，如果在转发表中找不到匹配的转发项，将 MAC 帧从除接收该 MAC 帧端口以外的所有其他端口发送出去，这种转发操作称为广播。如果在转发表中找到目的 MAC 地址匹配的转发项，且转发项中的转发端口和接收该 MAC 帧的端口不同，将 MAC 帧从转发项指定的转发端口发送出去。

为了实现冲突域之间的 MAC 帧过滤操作，即防止同一冲突域内终端之间传输的 MAC 帧转发到其他冲突域。如果在转发表中找到目的 MAC 地址匹配的转发项，且转发项中的转发端口和接收该 MAC 帧的端口相同，则丢弃该 MAC 帧。

网桥对转发表中的每一项转发项都设置了一个定时器，如果在规定时间内没有接收到以该转发项中 MAC 地址为源 MAC 地址的 MAC 帧，将从转发表中删除该项，这样做的原因在于网络中终端的位置不是一成不变的，因此，网桥端口和终端之间的关联必须是动态的。

4. 网桥无限扩展以太网

中继器或集线器是传输媒体扩展设备。由于电信号在传播过程中会发生衰减，因此，单段传输媒体的长度受到严格限制，如单段双绞线的长度必须小于 100m。中继器的信号再生功能使得由中继器互连的传输媒体长度得到扩展，这也是称中继器为物理层互连设备的主要原因。如果单从电信号传播质量考虑，两个终端之间串接的集线器可以有无穷个，但由于存在冲突域直径限制，使得两个终端之间串接的集线器数目受到严格限制。扩展以太网的另一种方法是用网桥互连多个冲突域，虽然每一个冲突域受冲突域直径限制，但由于网桥的互连级数没有限制，因此，经网桥扩展后的以太网的端到端传输距离可以无限大。

5. 全双工通信扩展无中继传输距离

图 1.15 是一个双端口网桥，如果是一个多端口网桥，而且每一个端口只连接一个终端，是否就不存在冲突域了？答案显然是否定的，仍然存在冲突域，只是每一个冲突域由终端和网桥端口组成。这种情况下，网桥端口和终端之间允许的最大物理距离就是转换成距离后的冲突域直径。假定网桥端口和终端网卡都符合 1000Mb/s 以太网标准，在最短帧长只有 64B 的情况下，它们之间允许的最大距离等于 $(512/(2\times1000\times10^6))\times(2/3)c=51.2\text{m}$。

1000Mb/s 传输速率、以双绞线为传输媒体的以太网采用 4 对双绞线进行传输，其中一对用于发送数据，一对用于接收数据，如果只有一个终端和网桥端口互连，可以采用全双工通信方式，这样，网桥端口和终端可以同时发送、接收数据，不再存在冲突域，也没有了争用总线的问题。对于光纤这一传输媒体，由于存在发送和接收光纤，也支持全双工通信方式，也可消除冲突域。因此可以得出如下结论：如果网桥每一个端口只连接一个终端，且终端和网桥端口之间采用全双工通信方式，冲突域将不复存在，终端和网桥端口之间传输距离不再受冲突域直径限制。同样，互连网桥的物理链路也可采用全双工通信方式，以此消除冲突域直径对两个网桥之间传输距离的限制。

6. 透明网桥的含义

透明是指虽然某种物质是存在的,但无法感觉到。将具有上述工作原理的网桥称为透明网桥,是因为某个终端向连接在同一冲突域内的另一个终端发送 MAC 帧的过程,与向连接在不同冲突域的另一终端发送 MAC 帧的过程是相同的,MAC 帧传输过程中是否经过网桥对源和目的终端是透明的。

7. 网桥工作过程举例

【例 1.4】 如果图 1.18 中的互连设备是集线器,计算出图 1.18 中的冲突域数量。

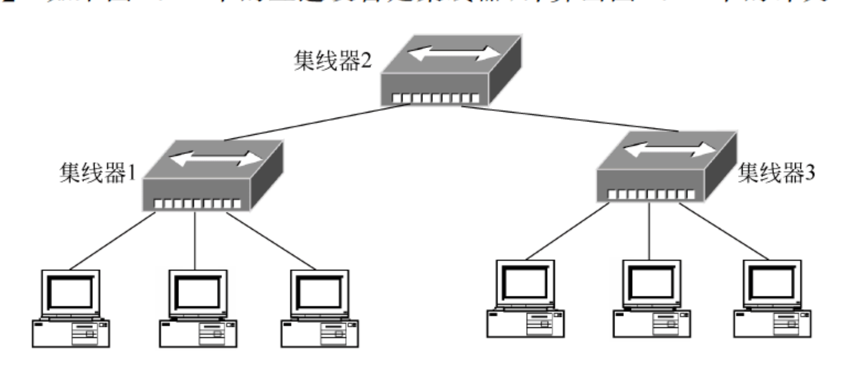


图 1.18 用集线器构建网络

【解析】 由于图中三个互连设备都是集线器,因此,整个网络是一个冲突域,所以图 1.18 中只有一个冲突域,根据以太网标准,图 1.18 所示网络结构只适用于 10Mb/s 以太网。

【例 1.5】 如果图 1.18 是用一个网桥互连两个集线器的网络结构(用网桥取代集线器 2),重新计算出图 1.18 中的冲突域数量。

【解析】 这种情况下,图 1.18 中的冲突域数量为 2,每一个冲突域范围是网桥端口加上集线器所连接的终端。

【例 1.6】 如果图 1.18 是用 3 个网桥构建的网络结构,每一个网桥端口只连接一个终端,终端和网桥之间、网桥和网桥之间采用全双工通信方式,重新计算出图 1.18 中的冲突域数量。

【解析】 这种情况下,图 1.18 中不存在冲突域,所有对冲突域的限制对上述假定下图 1.18 所示的网络结构不起作用。

【例 1.7】 如果图 1.19 是用 3 个网桥构建的网络结构,每一个网桥端口只连接一个终端,假定这三个网桥的初始转发表为空表,请给出按照顺序进行的终端 A→终端 D、终端 C→终端 D、终端 E→终端 A、终端 C→终端 E 的数据传输过程。

【解析】 当终端 A 需要给终端 D 发送数据时,它将需要发送的数据封装在 MAC 帧中的数据字段,以终端 A 的 MAC 地址(MAC A)为 MAC 帧的源 MAC 地址,终端 D 的 MAC 地址(MAC D)为 MAC 帧的目的 MAC 地址,然后将 MAC 帧发送给网桥 1,网桥 1 从端口 1 接收到该 MAC 帧,用该 MAC 帧携带的目的 MAC 地址去查找转发表,由于转发表为空,当然找不到匹配项,网桥 1 将该 MAC 帧从除接收端口(端口 1)以外的所有其他端口(端口 2、端口 3、端口 4)发送出去。用该 MAC 帧携带的源 MAC 地址去查找转发表,由于找不到匹配项,在转发表中添加一项,其中 MAC 地址为该 MAC 帧携带的源 MAC 地址,转发端口为

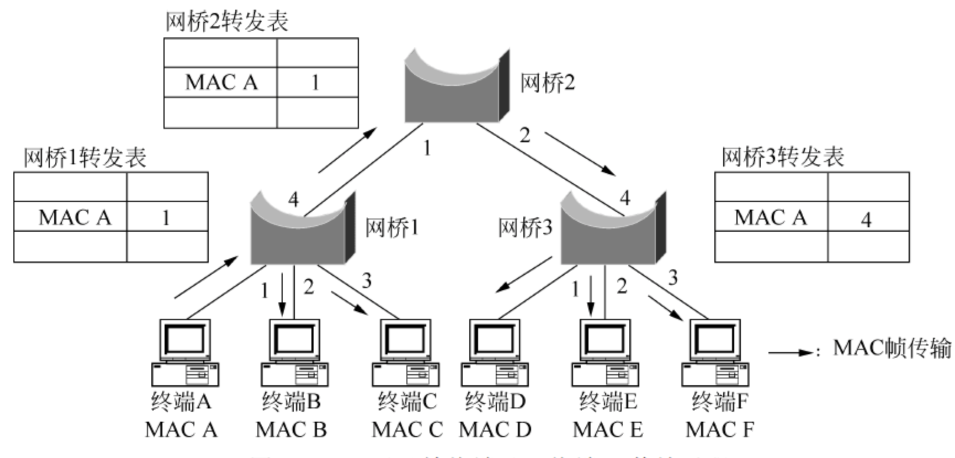


图 1.19　MAC 帧终端 A→终端 D 传输过程

接收该 MAC 帧的端口(端口 1),如图 1.19 所示。网桥 1 的广播操作使网桥 2 的端口 1 和终端 B、终端 C 均接收到该 MAC 帧,终端 B、终端 C 发现该 MAC 帧携带的目的 MAC 地址和自身的 MAC 地址不符,将该 MAC 帧丢弃。而网桥 2 和网桥 1 一样,由于在转发表中找不到该 MAC 帧对应的转发端口,广播该帧,并根据该 MAC 帧携带的源 MAC 地址在转发表中添加一项,如图 1.19 所示。

该 MAC 帧到达网桥 3,最终从网桥 3 的端口 1、端口 2 和端口 3(除接收该 MAC 帧的端口 4 以外的所有其他端口)转发出去,并根据该 MAC 帧携带的源 MAC 地址在网桥 3 的转发表中添加一项,如图 1.19 所示。从网桥 3 的端口 1、端口 2 和端口 3 转发出去的 MAC 帧到达终端 D、终端 E、终端 F,由于终端 D 自身的 MAC 地址(MAC D)和该 MAC 帧携带的目的 MAC 地址相同,继续处理该 MAC 帧,其他终端丢弃该 MAC 帧,传输过程结束。

终端 C→终端 D 的传输过程和终端 A→终端 D 的传输过程基本相同,由于终端 C 的 MAC 地址在转发表中找不到匹配项,因此,网桥 1、网桥 2 和网桥 3 的转发表中都增加了 MAC 地址为 MAC C 的项,如图 1.20 所示。

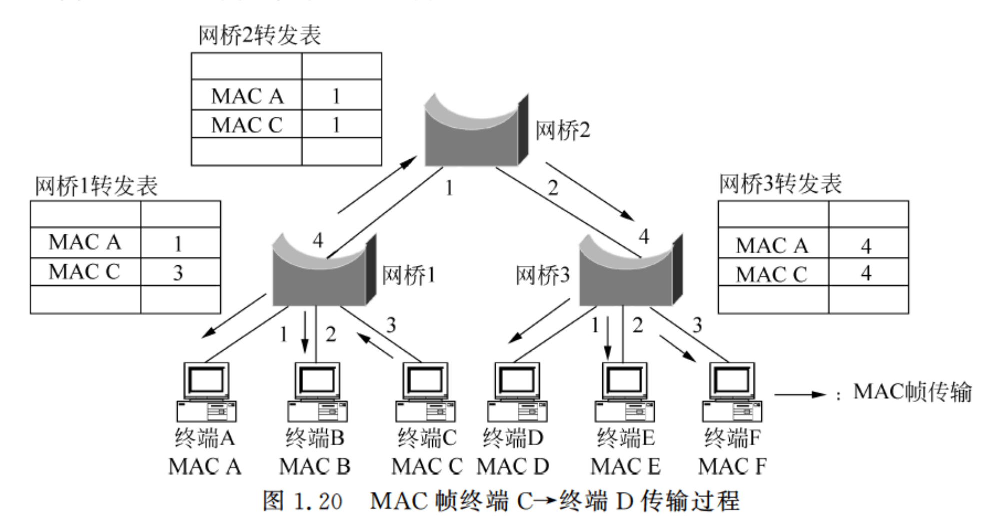


图 1.20　MAC 帧终端 C→终端 D 传输过程

终端 E→终端 A 传输过程如图 1.21 所示,当网桥 3 接收到终端 E 发送的源 MAC 地址为 MAC E,目的 MAC 地址为 MAC A 的 MAC 帧时,网桥 3 用该 MAC 帧携带的目的

MAC 地址去查找转发表，找到匹配项，并获知该 MAC 帧的转发端口为端口 4，将该 MAC 帧从端口 4 发送出去(不广播，只从端口 4 转发该 MAC 帧)，同时，在转发表中添加一项，该项的 MAC 地址为 MAC E(该 MAC 帧携带的源 MAC 地址)，转发端口为端口 2(网桥 3 接收到该 MAC 帧的端口)。

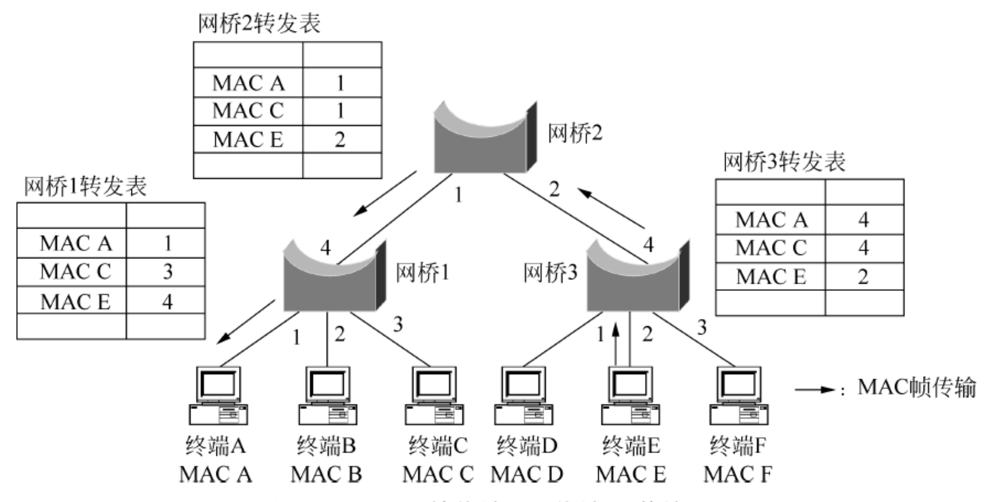


图 1.21　MAC 帧终端 E→终端 A 传输过程

从网桥 3 端口 4 发送出去的该 MAC 帧到达网桥 2 端口 2，网桥 2 同样根据转发表将该 MAC 帧从端口 1 发送出去，并在转发表中添加一项。网桥 1 依此操作，将该 MAC 帧通过端口 1 发送给终端 A，同时在转发表中添加一项，如图 1.21 所示。

MAC 帧从终端 C 传输到终端 E 的过程，和 MAC 帧从终端 E 传输到终端 A 的过程基本相同，由于网桥 1、网桥 2 和网桥 3 都能在转发表中找到和该 MAC 帧携带的目的 MAC 地址(MAC E)匹配的项，都能从指定端口发送该 MAC 帧，但由于转发表中已经存在和该 MAC 帧源 MAC 地址匹配的项，且该项给出的转发端口和网桥接收该 MAC 帧的端口相同，因此，网桥只刷新和转发表中该转发项关联的定时器(重新开始计时)，而不用添加新的项。MAC 帧传输过程如图 1.22 所示。

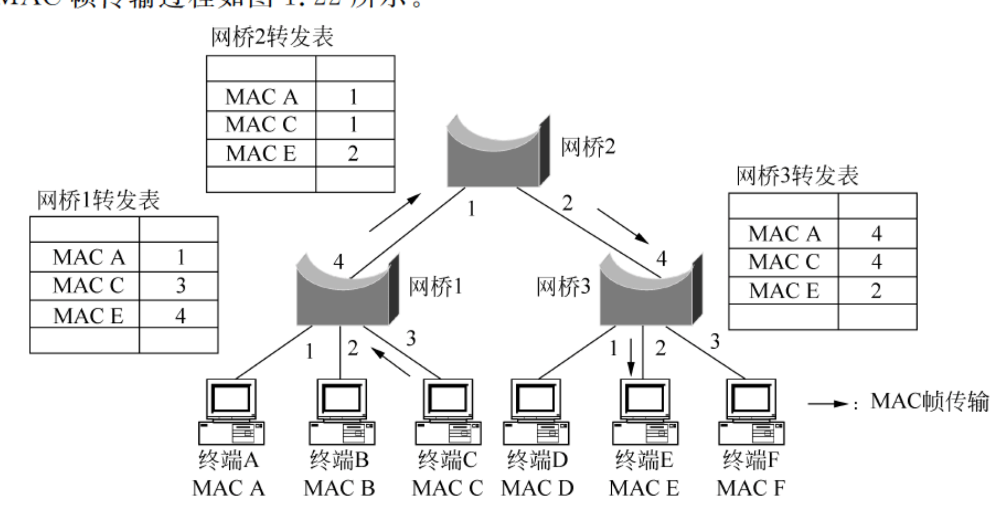


图 1.22　MAC 帧终端 C→终端 E 传输过程

当转发表中不存在和需要转发的 MAC 帧携带的目的 MAC 地址匹配的项时，网桥就广播该 MAC 帧，因此，在转发表完全建立之前，大量 MAC 帧是以广播方式传输的。为了使网络中所有终端在所有网桥的转发表中都有匹配项，每一个终端在加电启动后以自身 MAC 地址为源 MAC 地址，以广播地址(48 位全 1)为目的 MAC 地址广播一帧 MAC 帧，以便让网络中的所有网桥都在转发表中添加与该终端自身 MAC 地址匹配的项，这样，当有其他终端向该终端发送 MAC 帧时，该 MAC 帧经过的网桥就不需要以广播方式转发该 MAC 帧了。

1.3 交换机转发方式和交换机结构

1.3.1 交换机转发方式

交换机的基本工作原理和透明网桥是相同的，从本质上说，交换机是透明网桥的市场名称，但交换机在不同时期，为迎合市场需求，作了一些改进和改变。透明网桥是分组交换设备，采取存储转发方式，但交换机为了实现快速转发，采用了多种不同的转发方式。

1. 直通转发方式

交换机从输入端口开始接收 MAC 帧的第一位，到输出端口开始发送该 MAC 帧的第一位所需时间称为转发时延，为了减少转发时延，有的交换机采用直通转发方式(也称直接交换方式)，输入端口无须接收完整的 MAC 帧，在接收完 6 字节的目的地址字段后，开始进行 MAC 帧输入端口至输出端口的交换操作，并通过输出端口发送该 MAC 帧。直通转发方式能够有效减少转发时延。采用直通转发方式的前提是：①输入端口和输出端口的数据传输速率相同，②输出端口连接的是全双工信道且输出端口空闲。直通转发方式的缺陷是有可能转发长度小于 64B 的 MAC 帧，和已经发生传输错误的 MAC 帧。

2. 碎片避免转发方式

一旦检测到冲突发生，发送端将立即停止 MAC 帧发送，并发送 4 字节或 6 字节长度的阻塞信号(也称干扰信号)，因此，长度小于 64B 的 MAC 帧往往是因为发生冲突而产生的碎片(不完整 MAC 帧)，交换机转发这种类型的 MAC 帧会严重浪费链路带宽和交换机、终端的处理能力。为了避免转发碎片，交换机在接收到 64B 后，才开始 MAC 帧输入端口至输出端口的交换操作，并通过输出端口发送该 MAC 帧。这种转发方式称为碎片避免转发方式，除了不再转发碎片，碎片避免转发方式的前提和缺陷与直通转发方式相同。

3. 存储转发方式

存储转发方式下，交换机完整接收 MAC 帧，根据 MAC 帧中除 FCS 字段外的各个字段计算 CRC 码，并用计算出的 CRC 码和 MAC 帧中的 FCS 字段值比较，如果相等，表示 MAC 帧经过信道传输没有发生错误，如果不相等，表示 MAC 帧经过信道传输发生错误，交换机丢弃该 MAC 帧。只对没有检测出传输错误的 MAC 帧进行输入端口至输出端口的交换操作，并通过输出端口发送该 MAC 帧。

4. 三种转发方式比较

1）三种转发方式的转发时延

转发时延由三部分组成：一是交换机接收 MAC 帧中要求接收的字节长度所需要的时间，该时间取决于 MAC 帧中要求接收的字节长度和交换机输入端口的速率；二是交换机完成检错（只有存储转发方式具有），并根据 MAC 帧所携带的目的 MAC 地址确定输出端口的时间；三是在发生拥塞的情况（多个端口输入的 MAC 帧需要从同一个端口输出）下，在输出队列中排队等待的时间。为了方便比较，在计算转发时延时，只考虑输入端口接收 MAC 帧中要求接收的字节长度所需要的时间，忽略其他所需时间。

假定 MAC 帧的长度为 1518B，端口数据传输速率为 10Mb/s 的情况下，算出采用存储转发方式时的转发时延$=(1518\times8)/10^7=1.2144\times10^{-3}$s。

直通转发方式的转发时延与 MAC 帧长度无关，采用直通转发方式时的转发时延$=(6\times8)/10^7=4.8\times10^{-6}$s。

碎片避免转发方式的转发时延也与 MAC 帧长度无关，采用碎片避免转发方式时的转发时延$=(64\times8)/10^7=5.12\times10^{-5}$s。

2）结论

在转发时延方面，直通转发方式带有明显的优势，但存储转发方式的转发时延与交换机端口的传输速率和 MAC 帧长度有关，因此，直通转发方式在早期 10Mb/s 以太网交换机中作为一种改进交换机性能的重要技术予以推出。随着端口传输速率的提高，存储转发方式完整接收 MAC 帧所需的时间降低，而直通转发方式只有在特殊情况下才能实施，且取消了 MAC 层的差错控制功能，因此，对于目前 100Mb/s 以上传输速率端口的以太网交换机，尤其是多种传输速率端口并存的交换机，通常采用存储转发方式。

1.3.2　交换机结构

1. 交换机一般结构

交换机一般结构如图 1.23 所示，每一个交换机端口用于连接传输媒体，传输媒体可以是非屏蔽双绞线和光纤，存储转发方式下，由输入端口完成帧定界，即从通过传输媒体接收到的电信号或光信号序列中分解出每一帧 MAC 帧。完成对 MAC 帧的检错，丢弃传输出错的 MAC 帧，将没有出错的 MAC 帧存储到输入队列，根据 MAC 帧的目的 MAC 地址和转发表（也称 MAC 地址表）确定输出端口，通过交换结构将 MAC 帧从输入端口的输入队列交换到输出端口的输出队列，如果输出端口空闲，立即将 MAC 帧从输出端口发送出去，如果输出端口正在发送其他 MAC 帧，该 MAC 帧将在输出端口的输出队列中排队等候。直通转发方式下，输入端口接收完 6B 目的 MAC 地址，就开始根据 MAC 帧的目的 MAC 地址和转发表确定输出端口的过程。

交换机性能取决于以下两个性能参数：一是根据 MAC 帧的目的 MAC 地址和转发表确定输出端口所需的时间；二是 MAC 帧从输入端口交换到输出端口所需的时间。目前主要通过采用内容寻址存储器（Content Addressable Memory，CAM）来提高前一个性能参数，CAM 是一种以目的 MAC 地址作为地址输入，以转发端口作为内容输出的存储器，但这

种存储器的存储单元数量远远少于 2^{48} 个，因此，需要采用特殊的存储器结构和 MAC 地址至存储器地址的映射算法。提高后一种性能参数需要改进交换结构，目前常用的交换结构有共享总线和交叉矩阵，交叉矩阵的交换性能好于共享总线，但硬件结构也相对复杂。

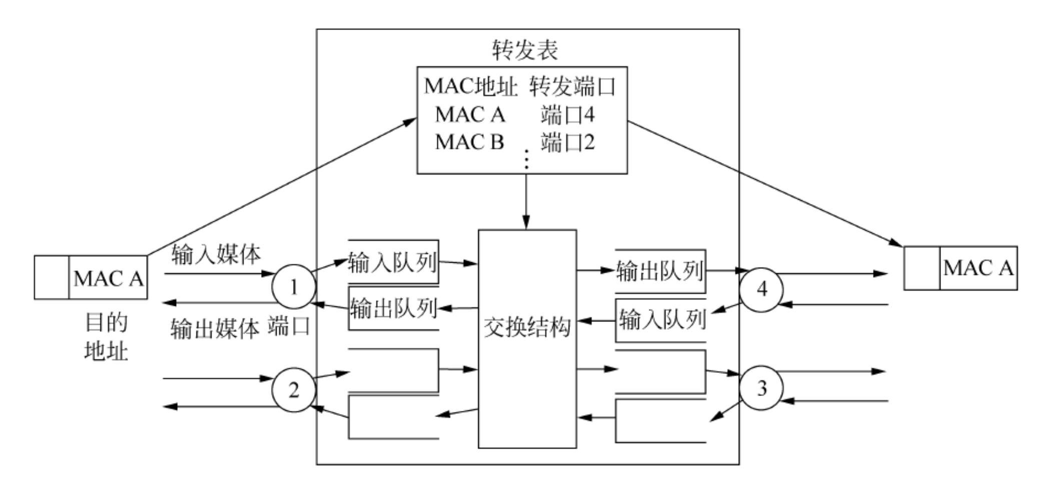


图 1.23　交换机一般结构

2. 共享总线交换结构

共享总线交换结构如图 1.24 所示，管理器和所有端口连接三组总线上，三组总线分别为数据总线(DB)、控制总线(CB)和结果总线(RB)，管理器负责总线仲裁和根据目的 MAC 地址和转发表(MAC 地址表)确定输出端口的功能。下面以终端 A 向终端 B 发送 MAC 帧为例讨论共享总线交换结构完成 MAC 帧从输入端口交换到输出端口的过程。

(1) 端口 1 完整接收 MAC 帧，完成对 MAC 帧检错，将没有传输错误的 MAC 帧放入输入队列，由总线控制器通过控制总线(CB)向管理器发出请求使用数据总线(DB)的信号。

(2) 如果数据总线空闲，管理器通过控制总线向端口 1 总线控制器发送允许使用数据总线信号。

(3) 端口 1 总线控制器通过数据总线发送 MAC 帧和控制信息，控制信息是除 MAC 帧外管理器完成地址学习、确定输出端口所需的全部信息，这里主要是输入端口号，以后还需要包括输入端口所属 VLAN 的 VLAN 标识符等。MAC 帧和控制信息被连接在数据总线上的所有端口和管理器接收，并存储。

(4) 管理器根据接收到的 MAC 帧的目的 MAC 地址和创建的 MAC 地址表确定输出端口，通过结果总线(RB)发送输出端口号。同时根据控制信息完成地址学习过程。

(5) 所有端口的总线控制器接收到输出端口号后，和自己保存的端口号比较，如果相同，将 MAC 帧放入输出队列，如果不同，丢弃该 MAC 帧。输出端口逐个输出存储在输出队列中的 MAC 帧。

共享总线交换结构任何时候只能实现两个端口之间 MAC 帧的单向传输，由于管理器根据接收到的 MAC 帧的目的 MAC 地址和创建的 MAC 地址表确定输出端口需要时间，为提高数据总线的利用率，将 MAC 帧经过数据总线传输过程与管理器确定 MAC 帧输出端口过程分为流水线上的两个操作步骤，在管理器确定前一帧 MAC 帧的输出端口的同时，允

许数据总线传输其他终端之间的 MAC 帧。

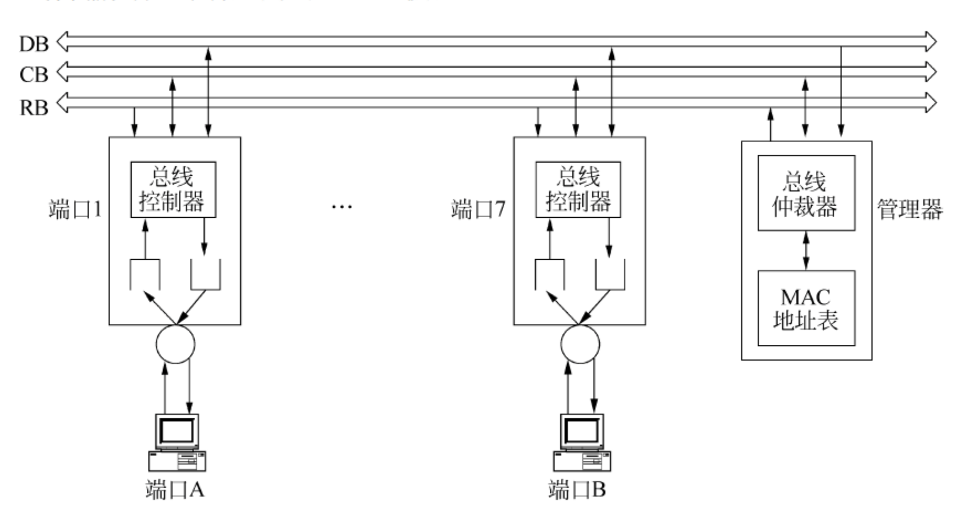


图 1.24 共享总线交换结构

3. 交叉矩阵交换结构

交叉矩阵能够同时建立不同端口对之间的双向传输通路，允许多对端口之间同时进行双向 MAC 帧传输，如图 1.25(a)所示。如果交换机有 N 个 Mb/s 传输速率的端口，理想的交叉矩阵的交换容量为 $2\times N\times$Mb/s。图 1.25(b)是交叉矩阵的一种实现，8 个交换机端口同时连接在 8 条横线和 8 条竖线上，横线和竖线之间存在开关，一旦开关闭合，横线和竖线相连，一旦开关断开，横线和竖线断开，图中黑点标识闭合的横线和竖线之间开关，图 1.25(b)中横线和竖线相连情况对应图 1.25(a)所示的多对端口之间的连接。需要强调的是，图 1.25(b)所示的是交叉矩阵原始实现方式，并不是目前交换机中采用的交叉矩阵的实现方式。

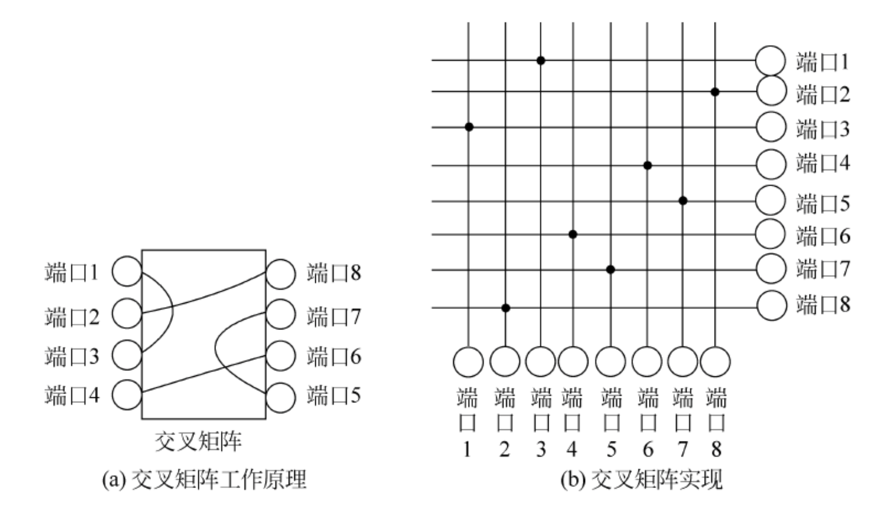


图 1.25 交叉矩阵

交叉矩阵交换结构如图 1.26 所示，所有端口和交叉矩阵相连，交叉矩阵可以同时在不同端口对之间建立双向传输通路，以此实现不同端口对之间 MAC 帧的并行传输。下面以

终端 A 向终端 B 发送 MAC 帧为例讨论交叉矩阵交换结构完成 MAC 帧从输入端口交换到输出端口的过程。

(1) 端口 1 完整接收 MAC 帧,完成对 MAC 帧检错,将没有传输错误的 MAC 帧放入输入队列,由总线控制器通过控制总线(CB)向管理器发送请求使用数据总线(DB)的信号。

(2) 如果数据总线空闲,管理器通过控制总线向端口 1 总线控制器发送允许使用数据总线信号。

(3) 端口 1 总线控制器通过数据总线发送控制信息,控制信息是管理器完成地址学习、确定输出端口所需的全部信息,这里主要是输入端口号、MAC 帧源和目的 MAC 地址等,以后还需要包括输入端口所属 VLAN 的 VLAN 标识符等,管理器接收、并存储控制信息。

(4) 管理器根据接收到的控制信息和创建的 MAC 地址表确定输出端口,通过结果总线(RB)发送输出端口号。同时根据控制信息完成地址学习过程。

(5) 端口 1 总线控制器接收到输出端口号后,生成用于要求交叉矩阵建立输入端口和输出端口之间双向传输通路的指令,并将指令发送给交叉矩阵,随后,将 MAC 帧发送给交叉矩阵,交叉矩阵通过已经建立的输入端口和输出端口之间双向传输通路,将 MAC 帧传输给输出端口。

(6) 端口 7 接收到 MAC 帧后,将 MAC 帧放入输出队列。输出端口逐个输出存储在输出队列中的 MAC 帧。

交叉矩阵交换结构和共享总线交换结构最大不同在于输入端口通过数据总线传输的仅仅是几十字节长度的控制信息,因此,控制信息传输和输出端口确定所需的时间较短。由于交叉矩阵能够同时建立不同端口对之间的双向传输通路,多对端口之间允许同时传输 MAC 帧。

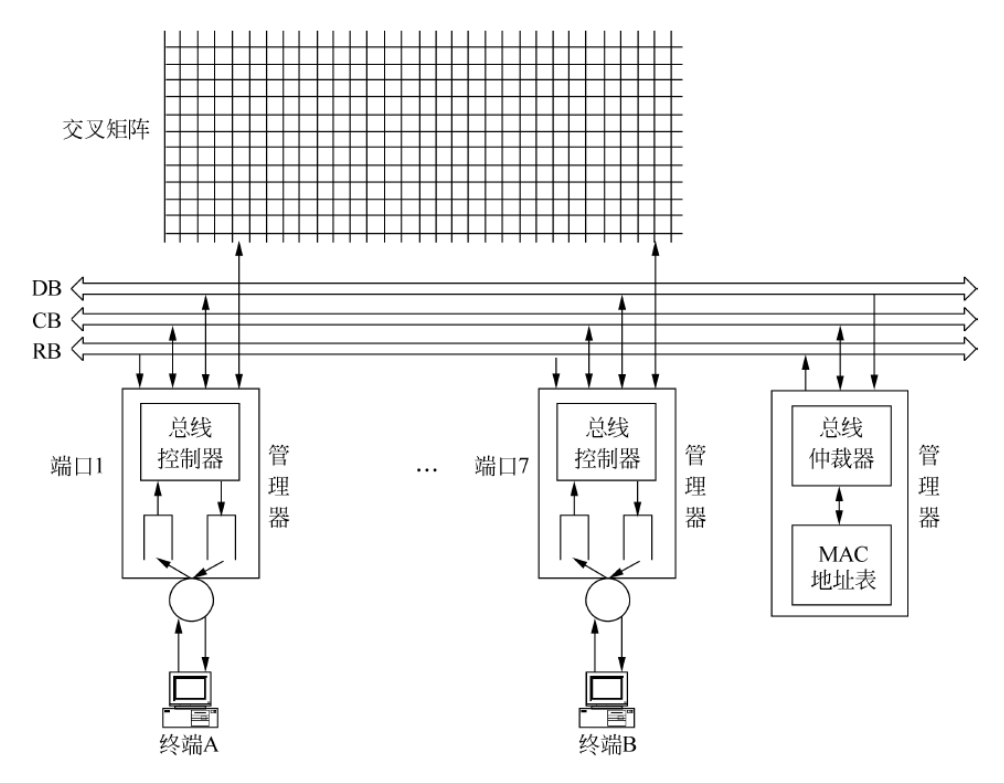


图 1.26 交叉矩阵交换结构

4. 交换式以太网的本质含义

交换机本质上是一个数据报分组交换机，如图 1.23 所示，转发表中各转发项用于指出通往由 MAC 地址指定的目的终端的传输路径。交换机有多个端口，每一个端口可以连接点对点信道，也可以连接广播信道。广播信道可以是单段总线，或是由中继器互连的多段总线，也可以是由集线器互连的多对双绞线缆。目前常见的广播信道是由集线器构成的星型冲突域。如果某个端口连接的是全双工点对点信道，则可以直接通过输出物理链路输出 MAC 帧，不需在输出 MAC 帧时进行 CSMA/CD 操作。如果某个端口连接的是广播信道或是半双工点对点信道，则通过 CSMA/CD 操作输出 MAC 帧。由于广播信道和半双工点对点信道本身是一个冲突域，所以最远距离受冲突域直径限制。

交换机在 MAC 帧端到端传输过程中完成两个功能：一是检测 MAC 帧经过每一段物理链路传输后是否出错，并丢弃出错的 MAC 帧；二是选择通往目的终端的传输路径。因此，图 1.27 所示的由交换机互连点对点信道或广播信道构成的网络就是一个数据报分组交换网络。根据 OSI 网络体系结构所定义的功能，点对点信道或广播信道实现物理层要求的基带信号传输功能。全双工点对点信道的链路层功能相对简单，只是完成 MAC 帧封装、帧定界及检错等功能，广播信道或半双工点对点信道还需要通过 CSMA/CD 操作解决信道争用问题。交换机实现网络层要求的路由功能。

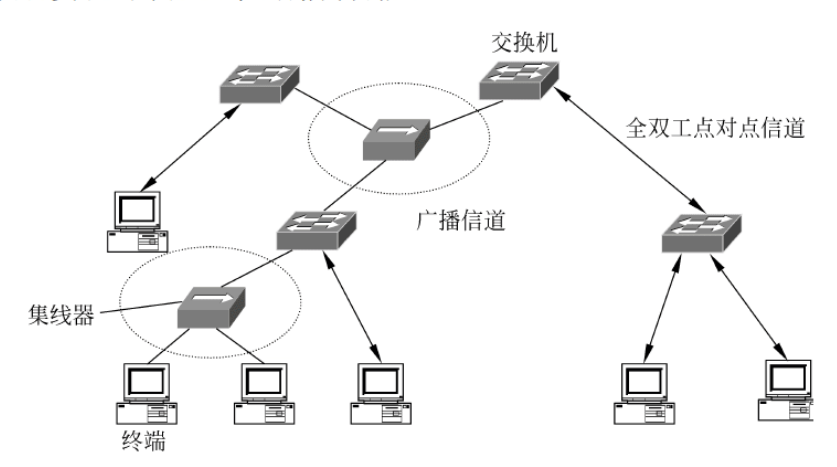


图 1.27　数据报分组交换网络

实际的讨论中为什么将交换机作为链路层设备？其主要原因是目前习惯将网际层等同于网络层，这样，网络层的功能被定义为路由 IP 分组，因此，只有用于互连不同类型传输网络的路由器被称为网络层设备，而交换机因为路由 MAC 帧被定义为链路层设备。这也表明 OSI 体系结构的功能定义适用于单种类型的传输网络，并不适用于互联的网络结构，因此，在以后的讨论中对设备按层分类的依据是该设备处理的对象，如果处理的对象是电信号或光信号，则为物理层设备。如果处理的对象是和特定传输网络相关的信息格式，如以太网的 MAC 帧，则为链路层设备。如果处理的对象是 IP 分组，则为网际层设备。为了和人们目前的习惯一致，也可以称为网络层设备。网络层设备也被称为三层设备，依次类推，链路层设备可以称为二层设备。以太网中由交换机和互连交换机的物理链路构成的端到端传输路径称为交换路径，以此凸显交换机的分组交换功能。

5. 例题解析

【例 1.8】 共享总线交换结构和交叉矩阵交换结构上连接 4 个端口，数据总线和端口连接交叉矩阵链路的带宽均为 1Gb/s，假定两对端口之间同时需要传输长度为 1000B 的 MAC 帧，控制信息长度为 32B，除完成控制信息和 MAC 帧传输需要时间外，其他操作所需时间忽略不计。求共享总线交换结构和交叉矩阵交换结构各自完成两对端口之间 MAC 帧传输所需时间。

【解析】 共享总线交换结构串行传输两组 MAC 帧和控制信息，所需时间 $=(2\times(1000+32)\times8)/(10^9)=1.6512\times10^{-5}$s。

交叉矩阵交换结构由于只需串行传输控制信息，两对终端之间可以并行传输 MAC 帧，因此，所需时间 $=((2\times32+1000)\times8)/(10^9)=8.512\times10^{-6}$s。

【例 1.9】 交换机连接终端和集线器方式如图 1.28 所示，假定终端后面的字符表示终端的 MAC 地址，初始转发表为空表，回答以下问题。

① 终端 A 发送的目的 MAC 地址为 B 的 MAC 帧到达哪些终端？

② 终端 B 发送的目的 MAC 地址为 A 的 MAC 帧到达哪些终端？

③ 终端 E 发送的目的 MAC 地址为 B 的 MAC 帧到达哪些终端？

④ 终端 B 发送的目的 MAC 地址为 E 的 MAC 帧到达哪些终端？

⑤ 终端 B 发送的目的 MAC 地址为广播地址的 MAC 帧到达哪些终端？

⑥ 终端 F 发送的目的 MAC 地址为 E 的 MAC 帧到达哪些终端？

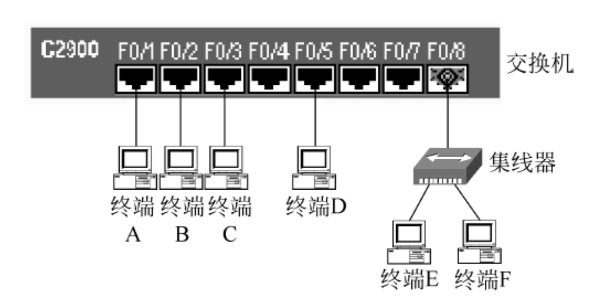


图 1.28 交换机连接终端和集线器方式

【解析】 ① 由于初始转发表为空表，转发表中没有与 B 的 MAC 地址匹配的转发项，该 MAC 帧被交换机广播，到达除终端 A 以外的所有其他终端(终端 B、终端 C、终端 D、终端 E 和终端 F)。转发表中增加 MAC 地址=A，输出端口(或转发端口)=F0/1 的转发项。

② 由于转发表中存在与 MAC 地址 A 匹配的转发项，交换机只从端口 F0/1 输出该 MAC 帧，该 MAC 帧只到达终端 A。转发表中增加 MAC 地址=B，输出端口(或转发端口)=F0/2 的转发项。

③ 由于连接终端 E 的设备是集线器，因此，该 MAC 帧被集线器广播，到达交换机端口 F0/8 和终端 F。由于转发表中存在与 MAC 地址 B 匹配的转发项，交换机从端口 F0/2 输出该 MAC 帧，该 MAC 帧到达终端 B。因此，该 MAC 帧到达终端 B 和终端 F。转发表中增加 MAC 地址=E，输出端口(或转发端口)=F0/8 的转发项。

④ 由于转发表中存在与 MAC 地址 E 匹配的转发项，交换机只从端口 F0/8 输出该

MAC 帧，由于连接端口 F0/8 的设备是集线器，被集线器广播的该 MAC 帧到达终端 E 和终端 F。

⑤ 由于 MAC 帧的目的地址是广播地址，该 MAC 帧被广播到除终端 B 以外的所有其他终端（终端 A、终端 C、终端 D、终端 E 和终端 F）。

⑥ 由于连接终端 F 的设备是集线器，该 MAC 帧被集线器广播，到达交换机端口 F0/8 和终端 E。由于转发表中存在与 MAC 地址 E 匹配的转发项，且输出端口 F0/8 是交换机接收该 MAC 帧的端口，交换机丢弃接收到的 MAC 帧。因此，该 MAC 帧只到达终端 E。

1.4　以太网标准

1.4.1　10Mb/s 以太网标准

1. 10BASE5

10BASE5 是用粗同轴电缆作为传输媒体的以太网标准，10 代表 10Mb/s，BASE 代表基带传输方式，即直接在电缆上传输数字信号，5 代表单段电缆的长度限制为 500m，超过 500m 需要由中继器互连的两段电缆组成，这个标准已经淘汰。

2. 10BASE2

10BASE2 是用细同轴电缆作为传输媒体的以太网标准，10 和 BASE 的含义和 10BASE5 相同，2 代表单段电缆的长度限制为 200m，超过 200m 需要由中继器互连的两段电缆组成，这个标准已经淘汰。

3. 10BASE-T

10BASE-T 是用双绞线作为传输媒体的以太网标准，它采用 4 对双绞线组成的双绞线电缆，用其中一对双绞线发送数据，另一对双绞线接收数据，因此，可以实现全双工通信。10BASE-T 的出现是以太网发展史上的一个里程碑，它同时引发了一个新的行业：综合布线，使得综合布线作为计算机网络的基础设施，在计算机网络的实施过程中成为必不可少的一部分。

10BASE-T 用于以集线器或以太网交换机为组网设备的以太网中，网络设备之间、网络设备和终端之间的距离必须小于 100m。

1.4.2　100Mb/s 以太网标准

1. 100BASE-TX

100BASE-TX 必须采用 5 类以上布线系统，和 10BASE-T 一样，它也只用于以集线器或以太网交换机为组网设备的以太网中，网络设备之间、网络设备和终端之间距离必须小于 100m。如果以集线器为组网设备，整个网络构成一个冲突域，冲突域直径必须小于 216m，这样，整个网络中最多只能有两个集线器级联。如果以以太网交换机为组网设备，由于以太

网交换机的互连级数不受限制，导致网络覆盖距离不受限制。如果以太网交换机之间、以太网交换机和终端之间均采用全双工通信方式，就可消除冲突域，无中继通信距离不再受冲突域直径限制。

支持 100BASE-TX 的以太网交换机端口或网卡一般都支持 10BASE-T，在标明速率时，用 100/10BASE-TX 表示同时支持 100BASE-TX 和 10BASE-T，而且能够根据对方端口或网卡的速率标准自动选择速率标准（如果对方支持 100BASE-TX，则选择 100BASE-TX，如果对方只支持 10BASE-T，则选择 10BASE-T）。

2. 100BASE-FX

用双绞线作为传输媒体有一些限制：一是距离较短，不要说楼宇之间，就是同一楼层两端之间的距离都有可能超出 100m；二是必须要避开强电和强磁设备；三是封闭性不够，不能用于室外。因此，室外通信或超过 100m 的室内通信均采用光缆，而且室外通信必须采用铠装光缆——一种封闭性很好又有金属支撑和保护的光缆，可直埋地下或架空。

100BASE-FX 采用两根 50/125μm 或 62.5/125μm 的多模光纤，分别用于发送和接收数据，因此，支持全双工通信方式。如果两个 100BASE-FX 端口（通常情况下，一个是以太网交换机端口，另一个是以太网交换机端口或网卡）以全双工方式进行通信，它们之间的传输距离可达 2km。但如果以半双工方式进行通信，传输距离在 500m 左右，这是由于一旦采用半双工通信方式，则两个 100BASE-FX 端口之间就构成一个冲突域，对于 100BASE-FX 而言，512 位的最短帧长将冲突域直径限制为 2.56μs，换算成物理距离，大约等于 $2/3c \times 2.56 \times 10^{-6} = 2 \times 10^{8} \times 2.56 \times 10^{-6} = 512m$。因此，光纤连接的两个端口之间只有在采用全双工通信方式的情况下，才能真正体现光纤传输的远距离特点。

1.4.3 1Gb/s 以太网标准

1. 1000BASE-T

1000BASE-T 必须采用 5e 类以上的布线系统，支持 1000BASE-T 标准的端口通常也支持 100BASE-TX 标准，因此，常常标记成 1000/100/10BASE-TX，而且能够根据双绞线另一端连接的端口所支持的速率标准，从高到低自动选择速率标准。

2. 1000BASE-SX

1000BASE-SX 在全双工通信方式（许多 1Gb/s 以太网光纤端口只支持全双工通信方式）下，如果采用 62.5/125μm 多模光纤，其传输距离可达 225m，如果采用 50/125μm 多模光纤，其传输距离可达 500m。

3. 1000BASE-LX

1000BASE-LX 在全双工通信方式下，采用 9μm 单模光纤，其最小传输距离为 2km，不同 1000BASE-LX 端口，由于采用的激光强度不一样，其传输距离可在 2km～70km 之间。

市场上，9μm 单模光纤价格比 62.5/125μm 多模光纤便宜，关键是 1000BASE-LX 端口的价格是 1000BASE-SX 端口的 2 倍，因此，目前采用 1000BASE-LX 的成本比较高。

1.4.4 10Gb/s 以太网标准

1. 10GBASE-LR

10GBASE-LR 只能工作在全双工通信方式，采用单模光纤作为传输媒体，传输距离为 10km。很显然，交换和全双工通信方式完全消除了冲突域直径问题，使得以太网无论在传输速率上，还是无中继传输距离上，都成为城域网(Metropolitan Area Network，MAN)的最佳选择之一。

2. 10GBASE-ER

10GBASE-ER 只能工作在全双工通信方式，采用单模光纤作为传输媒体，传输距离为 40km。

10Gb/s 以太网从 2004 年推向市场后，逐渐成为校园网主干网络所采用的技术，在城域网中也和同步数字体系(Synchronous Digital Hierarchy，SDH)并驾齐驱，随着 10Gb/s 以太网逐渐成为 LAN 和 MAN 主流技术，10GBASE-T 标准与 7 类布线系统的出台，10Gb/s 以太网也会像 1Gb/s 以太网一样得到普及。

习题

1.1 什么是网络拓扑结构？目前存在哪些以太网拓扑结构？

1.2 802.3 标准局域网和以太网有什么区别，目前使用的以太网是否是 802.3 标准局域网？为什么？

1.3 冲突域直径是如何确定的？限制冲突域直径的主要因素是信号衰减吗？

1.4 什么是帧定界？以太网如何实现帧定界？

1.5 以太网不采用出错重传的差错控制机制，只是在接收端对接收到的 MAC 帧进行差错检验，丢弃传输出错的 MAC 帧，这种简单的差错检验机制，对以太网提出了什么要求？

1.6 以太网最短帧长是如何确定的？为什么必须检测到任何情况下发生的冲突？

1.7 后退算法如何体现它的自适应性？

1.8 什么是捕获效应？总线型以太网适合传输类似数字语音数据这样的多媒体数据吗？为什么？

1.9 假定单根总线的长度为 1km，传输速率为 1Gb/s，信号传播速度为(2/3)c，求最短帧长。

1.10 10Mb/s 以太网中某个终端在检测到冲突后，后退算法选择了随机数 $r=100$。问该终端需要等待多长时间才能发送数据？如果是 100Mb/s 的以太网呢？

1.11 终端 A 和 B 在同一个 10Mb/s 以太网网段上，它们之间的传播时延为 225 比特时间，假定在时间 $t=0$ 时，终端 A 和 B 同时发送了数据帧，在 $t=225$ 比特时间时同时检测到冲突发生，并在 $t=225+48=273$ 比特时间内发送完干扰信号，假定终端 A 和 B 选择的随机数分别是 0 和 1，回答：

① 终端 A 和终端 B 何时重传数据帧。

② 终端 A 重传的数据何时到达终端 B。

③ 终端 A 和终端 B 重传的数据会不会再次发生冲突。

④ 终端 B 在后退延迟后是否立即重传数据帧。

1.12 有 10 个终端连接到以太网上,试计算以下三种情况下每一个终端分配到的平均带宽。

① 10 个终端连接到 10Mb/s 集线器。

② 10 个终端连接到 100Mb/s 集线器。

③ 10 个终端连接到 10Mb/s 以太网交换机。

1.13 假定终端 A、B、C 和 D 连接在总线型以太网上,当终端 D 传输数据帧时,终端 A、B 和 C 开始侦听总线,画出终端 A、B 和 C 完成数据帧传输的流程图,要求:

① 成功传输数据帧顺序为终端 B、C 和 A。

② 传输过程中至少发生 4 次冲突。

1.14 以太网上只有两个终端,它们同时发送数据,发生了冲突,于是按截断二进制指数类型后退算法进行重传,重传次数记为 i,$i=1,2,3,\cdots$,试计算第 1 次、第 2 次、第 3 次重传失败的概率以及某个终端成功发送数据之前的平均重传次数 L。

1.15 以太网传输速率从 10Mb/s 发展到 100Mb/s、1Gb/s、10Gb/s 的主要技术障碍是什么?如何解决?讨论一下以太网最终能够成为 LAN、MAN 主流技术的原因。

1.16 假定图 1.29 中作为总线的电缆中间没有接任何中继设备,MAC 帧的最短帧长为 512b,电信号在电缆中的传播速度为 2/3 c(c 为光速),分别计算出 10Mb/s、100Mb/s、1000Mb/s 以太网所允许的电缆最长距离。

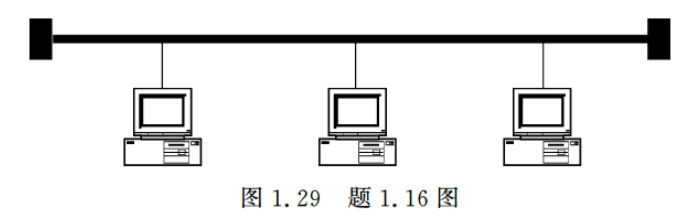


图 1.29 题 1.16 图

1.17 网桥分割冲突域的原理是什么?网桥如何实现属于不同冲突域的终端之间通信功能?

1.18 网桥是分组交换设备的依据是什么?

1.19 为什么说交换到无限?

1.20 为什么说交换式以太网是一个广播域?讨论一下广播带来的危害。

1.21 现有 5 个终端分别连接在三个局域网上,并且用两个网桥连接起来,如图 1.30 所示,每个网桥的两个端口号都标明在图上。开始时,两个网桥中的转发表都是空表,后来进行以下传输操作:H1→H5,H3→H2,H4→H3,H2→H1,试将每一次传输操作发生的有关事项填写在表 1.3 中。

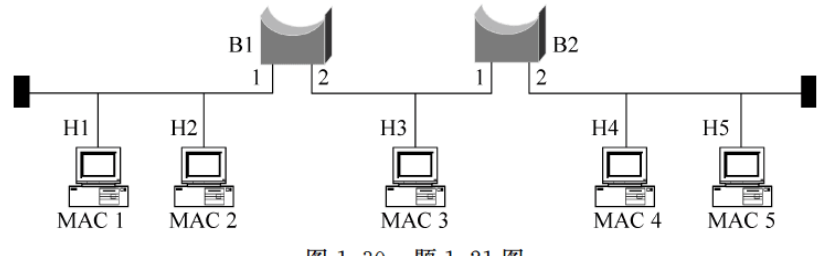


图 1.30 题 1.21 图

表 1.3　题 1.21 表

传输操作	网桥 1 转发表		网桥 2 转发表		网桥 1 的处理（转发、丢弃、登记）	网桥 2 的处理（转发、丢弃、登记）
	MAC 地址	转发端口	MAC 地址	转发端口		
H1→H5						
H3→H2						
H4→H3						
H2→H1						

1.22　图 1.31 所示网络结构有多少个冲突域？有多少个广播域？

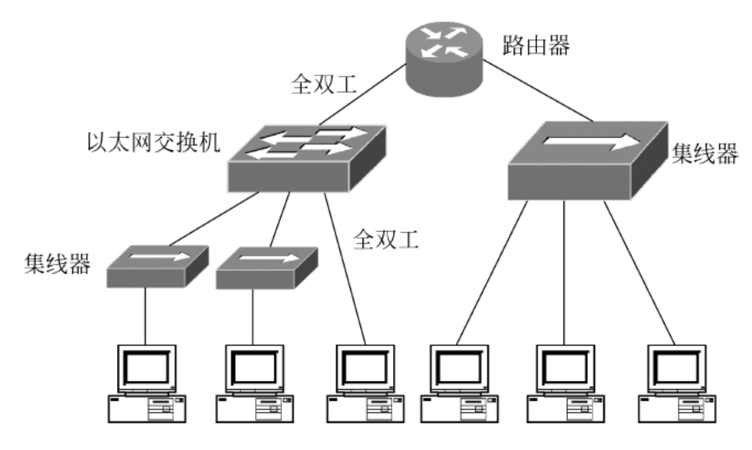


图 1.31　题 1.22 图

1.23　根据图 1.32 所示网络结构，假定所有以太网交换机的初始转发表为空表，给出完成终端 A→终端 B，终端 E→终端 F，终端 C→终端 A 数据帧传输后各个以太网交换机转发表内容。

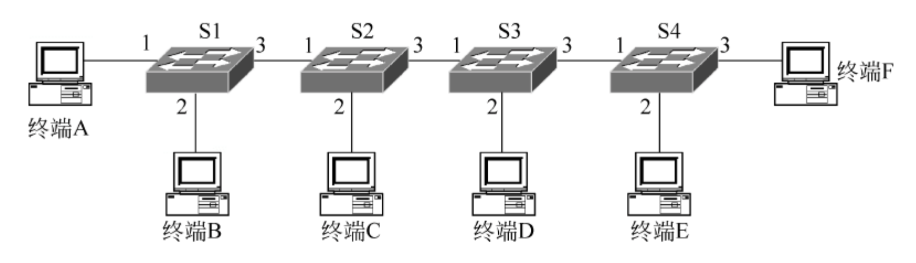


图 1.32　题 1.23 图

1.24　网络结构如图 1.33 所示，根据传输媒体为双绞线和光纤这两种情况，分别计算终端 A 和终端 B 之间的最大传输距离。假定集线器的信号处理时延为 0.56μs。

1.25　网络结构如图 1.34 所示，假定交换机初始转发表为空表，给出依次进行①～⑤ MAC 帧传输时，交换机 1 和交换机 2 完成的操作及转发表变化过程。

① 终端 A→终端 B。

② 终端 G→终端 H。

③ 终端 B→终端 A。

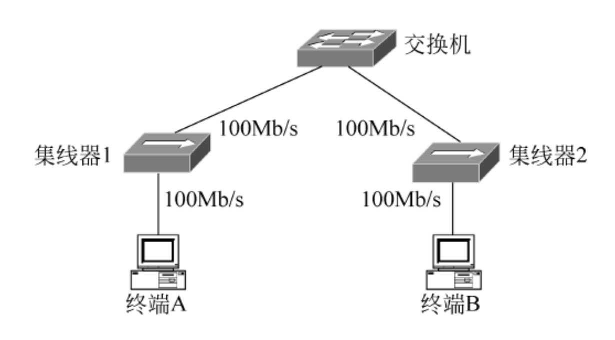


图 1.33　题 1.24 图

④ 终端 H→终端 G。

⑤ 终端 E→终端 H。

⑥ 如果将终端 A 移到交换机 1 端口 5 后，进行终端 E→终端 A 的 MAC 帧传输过程，会发生什么情况，如何解决？

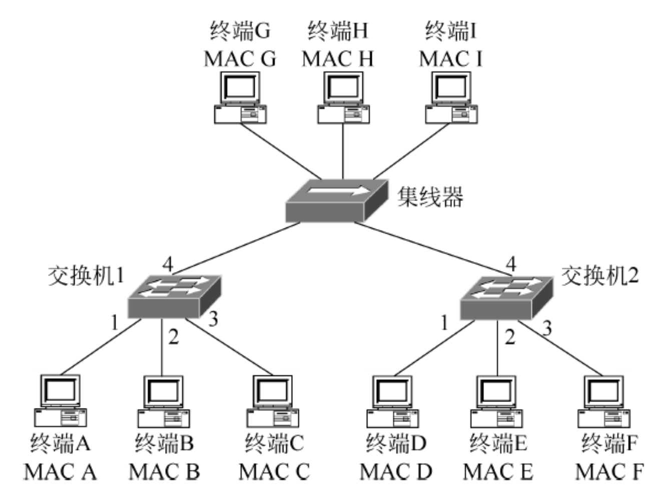


图 1.34　题 1.25 图

1.26　图 1.35 是连接某一幢楼内各个房间中终端的网络拓扑结构图，假定楼高为 30m，楼长为 90m，当图中设备的端口速率分别是 10Mb/s 和 100Mb/s 时，哪些设备可以是以太网交换机或集线器？哪些设备只能是以太网交换机？为什么？

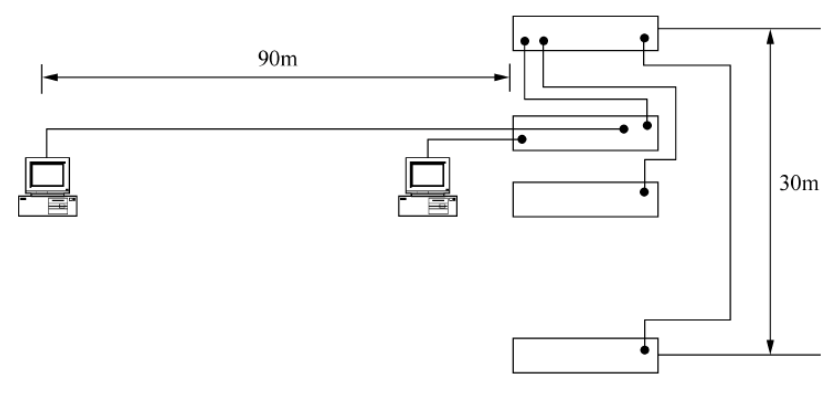


图 1.35　题 1.26 图

1.27　有两幢楼间距离超过500m的楼，每幢楼有5层，每层有20个房间，每个房间至少有一台终端，现在要求设计能够把所有房间中终端连接在一起的交换式以太网，请给出设备配置(多少端口、端口采用的以太网标准)，并说明原因。

1.28　共享总线交换结构和交叉矩阵交换结构上连接8个端口，数据总线和端口连接交叉矩阵链路的带宽均为1Gb/s，假定4对端口之间同时需要传输长度为1000B的MAC帧，控制信息长度为32B，除完成控制信息和MAC帧传输需要时间外，其他操作所需时间忽略不计。求共享总线交换结构和交叉矩阵交换结构各自完成两对端口之间MAC帧传输所需时间。

第2章 虚拟局域网

交换机和交换式以太网消除了冲突域直径与最短帧长之间的相互制约和共享式以太网的带宽瓶颈，实现了端到端传输路径的无限延长（交换到无限），全双工点对点链路使得交换机之间、交换机和终端之间的无中继传输距离只受传输媒体特性和信号质量的限制。但交换式以太网本身是一个广播域，大量高层协议通过广播实现数据传输，交换机转发 MAC 帧的操作过程也使得大量单播 MAC 帧以广播方式传输，这些以广播方式传输的 MAC 帧到达广播域内的每一个终端，这不仅浪费了链路带宽和终端的处理能力，还造成严重的安全隐患。虚拟局域网通过分割广播域解决了广播传输方式造成的这些问题。

2.1 广播域和广播传输方式

2.1.1 单播传输方式和广播传输方式

如果交换机从某个端口接收到目的地址为单播地址的 MAC 帧，且在转发表中找到 MAC 帧目的地址匹配的转发项，交换机只从转发项指定端口将 MAC 帧转发出去，这种转发方式称为单播传输方式。

如果交换机从某个端口接收到目的地址为广播地址的 MAC 帧，或者目的地址虽然是单播地址，但在转发表中找不到 MAC 帧目的地址匹配的转发项，交换机从除接收该 MAC 帧的端口以外的所有其他端口将 MAC 帧转发出去，这种转发方式称为广播传输方式。

用集线器或总线构成的以太网是一个共享式以太网，任何终端发送的 MAC 帧能够被其他所有终端接收，因此，在共享式以太网中，即在同一个冲突域中，单播传输方式和广播传输方式是相同的。但在由交换机构成的以太网中，对于单播方式传输的 MAC 帧，交换机通过查找转发表确定单一转发端口转发该 MAC 帧，网络中其他非目的终端接收不到该 MAC 帧。图 2.1 给出了共享式和交换式以太网转发终端 A→终端 B 的 MAC 帧的差别。

如果 MAC 帧的目的 MAC 地址为广播地址，或者虽然 MAC 帧的目的 MAC 地址为单播地址，但在交换机转发表中找不到和该 MAC 帧的目的 MAC 地址匹配的转发项，该 MAC 帧仍将广播到网络中的所有其他终端，如图 2.2 所示。

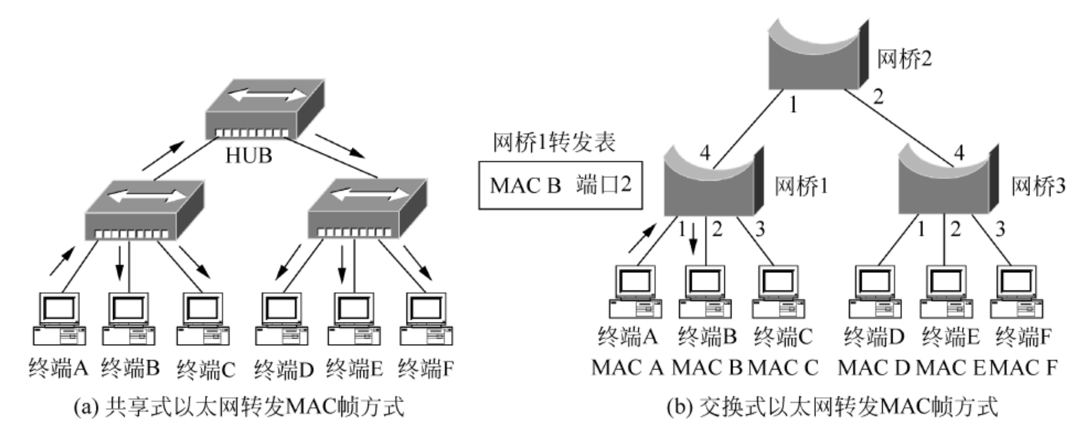


图 2.1 共享式和交换式以太网转发 MAC 帧的差别

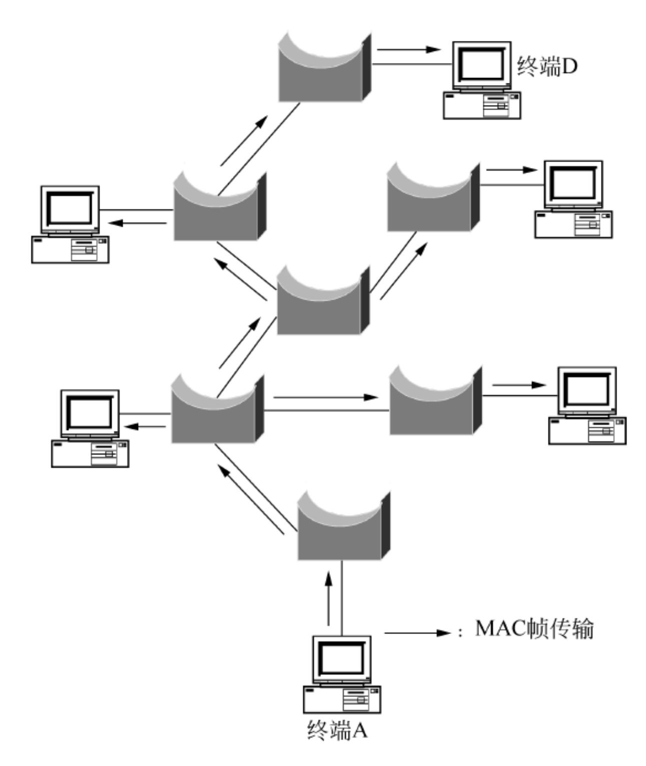


图 2.2 交换式以太网广播传输方式

2.1.2 广播域

将广播域定义为以目的地址为广播地址的广播帧在网络中的传播范围，并由此可以得出广播域和冲突域的最大区别在于任何终端发送的任何 MAC 帧均覆盖整个冲突域，而只有以广播方式传输的 MAC 帧才可能覆盖整个广播域。这种具有广播方式传输特性的网络称为广播型网络。虽然由交换机构建的交换式以太网消除了冲突域带来的问题，但整个交换式以太网仍然是一个广播域。在以太网中，广播操作是不可避免的，一是只有在不断的广播操作中，交换机才能建立起完整的转发表；二是 TCP/IP 协议栈中的许多协议如 ARP、DHCP 都是面向广播的协议。如果整个以太网就是一个广播域，而广播操作又频繁地进

行,网络带宽的利用率及终端的负荷都将成为问题。更为严重的是,由于广播传输方式将MAC帧传输给广播域中的每一个终端,将引发MAC帧中数据的安全性问题。

2.1.3 传统分割广播域的方式

为了解决广播引发的问题,只有将一个大型的交换式以太网分割成若干个较小的子网,用路由器将这些子网互连在一起,如图2.3所示。每一个子网就是一个广播域,即使是目的MAC地址为广播地址的MAC帧,也不能跨越路由器从一个子网广播到另一个子网。使用子网这个术语是为了说明这些小型以太网是划分大型以太网后产生的,实际上,每一个子网就是一个独立的以太网。

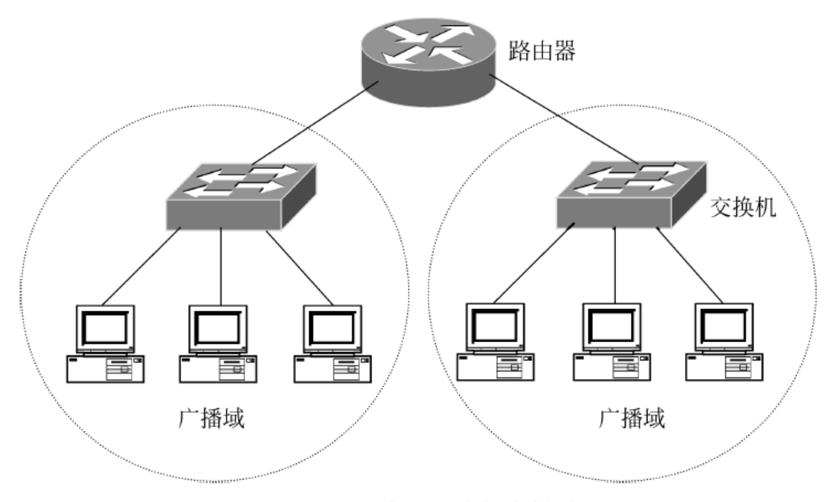


图2.3 用路由器分割广播域

图2.3是虚拟局域网(Virtual LAN,VLAN)出现前的一种常见网络拓扑结构,用以太网交换机(或网桥)构成若干较小的以太网,用路由器将这些小型以太网互连成一个大型网络。但这种结构存在一些缺陷:一是由于传输距离的限制,某个交换机所连接的终端必须局限在相对较小的地理范围内,导致子网必须以物理地域作为划分单位;二是一旦网络完成设计和实施,增加或删除一个子网,或者重新划分子网都是一件十分不容易的事。但在实际应用中,人们非常希望不受物理地域限制来划分子网,如一个课题组包含了数学系、计算机系和无线电系的若干教员,这些教员分散在不同的大楼内,但需要相互共享一些与课题有关的文件和程序,简单而安全的共享方式要求他们所使用的终端必须在一个子网内。还有,为了对不同应用的服务器设置不同的安全等级,也常常需要重新划分子网,将不同安全等级的服务器分配到相应子网中,但这种分配最好不需要对现有网络架构进行物理调整。

2.2 VLAN定义和分类

2.2.1 VLAN定义

真正解决广播引发的问题的方法必须做到:①可以在不改变一个大型交换式以太网的物理连接的前提下,任意划分子网;②每一个子网中的终端具有物理位置无关性,即每一个

子网可以包含位于任何物理位置的终端；③子网划分和子网中终端的组成可以通过配置改变，且这种改变对网络的物理连接不会提出任何新的要求。这就要求有一种全新的子网划分(或叫广播域分割)技术出现，这种技术就是虚拟局域网技术。

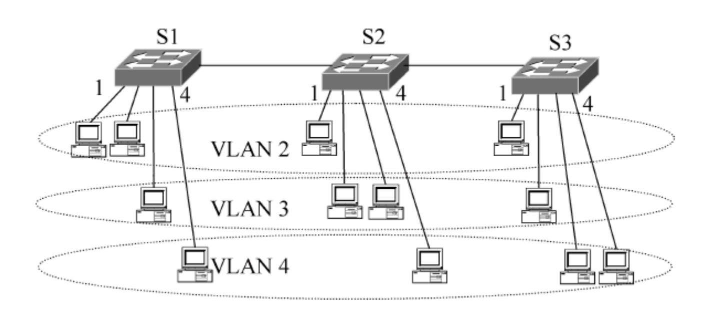


图 2.4　VLAN 原理图

图 2.4 所示是一个在物理交换式以太网上划分三个 VLAN 的实例，物理交换式以太网由 3 台交换机互连而成，每一个交换机通过端口 1～端口 4 连接 4 个终端，12 个终端被划分为 3 个 VLAN，每一个 VLAN 可以包含任意数量、位于任意物理位置的终端。通过配置，在不需要改变交换式以太网物理连接的前提下，可以任意改变 VLAN 数量和每一个 VLAN 包含的终端。对于图 2.4 所示的物理交换式以太网，可以通过配置，将 3 个 VLAN 变为 4 个 VLAN，增加的 VLAN 5 可以包含 12 个终端中的任意终端，但一般情况下，每一个终端只能属于一个 VLAN。

VLAN 完全等同于一个独立的交换式以太网。虽然，多个 VLAN 可以存在于同一个由交换机组成的物理交换式以太网中，但这些 VLAN 是相互独立的，属于不同 VLAN 的终端之间是不能相互通信的。为了讨论方便，将网桥作为一种无论物理上，还是逻辑上都只能属于单个 VLAN 的设备，而将以太网交换机作为一种支持 VLAN 划分的设备，一旦某台以太网交换机被划分为多个 VLAN，该以太网交换机等同于若干个功能独立的网桥。

2.2.2　VLAN 分类

为了实现广播域分割，必须能够将连接在物理交换式以太网上的终端按照用户制定的分配原则分配到各个 VLAN 中，根据将终端分配到 VLAN 的方式，可以将 VLAN 分为基于端口划分的 VLAN、基于终端 MAC 地址划分的 VLAN、基于协议划分的 VLAN 和基于终端网络地址划分的 VLAN。

1. 基于端口划分的 VLAN

基于端口划分的 VLAN 如图 2.5 所示，创建某个 VLAN，将交换机端口分配给某个 VLAN，建立端口和 VLAN 之间的绑定，每一个 VLAN 可以包含任意的交换机端口组合。对应图 2.5 所示的 VLAN 划分，建立表 2.1 所示的端口和 VLAN 之间的绑定。直接连接终端的交换机端口称为接入端口，一般情况下，每一个接入端口只能分配给一个 VLAN。

基于端口划分的 VLAN 中的基本成员是端口，根据表 2.1 所示的端口和 VLAN 之间的绑定，如果终端 A 接入端口 1，则终端 A 属于 VLAN 2，在 VLAN 2 中广播的 MAC 帧能够到达终端 A，但如果将终端 A 和终端 B 互换，即通过端口 2 接入终端 A，终端 A 将属于

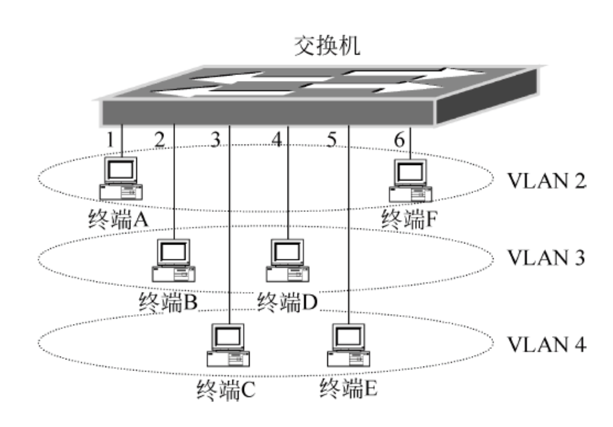


图 2.5 基于端口划分的 VLAN

VLAN 3。因此，基于端口划分的 VLAN 由端口组成，不是由终端组成，终端接入属于某个 VLAN 的端口后，才能确定该终端所属的 VLAN。基于端口划分的 VLAN 是最常见的 VLAN，端口和 VLAN 之间的绑定需要手工配置，如果需要改变某个 VLAN 中的端口组合，必须通过手工配置重新建立端口和 VLAN 之间的绑定。

表 2.1 端口和 VLAN 之间的绑定

端口	VLAN ID
端口 1	VLAN 2
端口 2	VLAN 3
端口 3	VLAN 4
端口 4	VLAN 3
端口 5	VLAN 4
端口 6	VLAN 2

2. 基于 MAC 地址划分的 VLAN

对于基于端口划分的 VLAN，如果要求将某个终端固定分配给某个 VLAN，即要求建立终端与 VLAN 之间的绑定，同时又允许该终端漫游，则必须在该终端可能漫游到的地方，留有分配给该 VLAN 的端口。这一方面可能造成交换机端口浪费，另一方面所有插入这些端口的终端都被作为该 VLAN 的成员，无法保证与该 VLAN 建立绑定的终端的唯一性。

为了建立终端与 VLAN 之间的绑定，必须建立终端标识符与 VLAN 之间的绑定，最常用作终端标识符的是 MAC 地址，因此，可以建立如表 2.2 所示的 MAC 地址与 VLAN 之间的绑定，交换机不是根据终端接入交换机的端口确定该终端属于的 VLAN，而是通过接收到的 MAC 帧的源 MAC 地址确定发送该 MAC 帧的终端所属的 VLAN。如果终端 A～终端 F 的 MAC 地址分别为 MAC A～MAC F，如果需要按照图 2.5 所示的 VLAN 组成将终端分配给各个 VLAN，建立如表 2.2 所示的 MAC 地址与 VLAN 之间的绑定。这种 VLAN 划分方式下，即使将终端 A 与终端 B 互换(即将终端 A 接入交换机端口 2，将终端 B 接入交换机端口 1)，终端 A 仍然属于 VLAN 2，终端 B 仍然属于 VLAN 3。

表 2.2 MAC 地址和 VLAN 之间的绑定

MAC 地址	VLAN ID
MAC A	VLAN 2
MAC B	VLAN 3
MAC C	VLAN 4
MAC D	VLAN 3
MAC E	VLAN 4
MAC F	VLAN 2

基于 MAC 地址划分的 VLAN 中的基本成员是终端，某个端口属于哪一个 VLAN，由接入该端口的终端的 MAC 地址确定。当终端 A 漫游，只要表 2.2 所示的 MAC 地址与 VLAN 之间的绑定不变，任何接入终端 A 的交换机端口都属于 VLAN 2，与该交换机端口的位置和编号无关。

3. 基于协议划分的 VLAN

基于协议划分的 VLAN 中的基本成员是终端，根据终端使用的网络协议来确定终端所属的 VLAN，为了确定某个终端所属的 VLAN，必须建立如表 2.3 所示的网络协议与 VLAN 之间的绑定，交换机根据接收到的分组的协议类型和网络协议与 VLAN 之间的绑定来确定发送分组的终端所属的 VLAN。由于目前基本使用 IP，因此，基于协议划分 VLAN 会使 IP 对应的广播域过大，失去划分 VLAN 的意义，因此，这种 VLAN 划分方式目前很少使用。

表 2.3 网络协议和 VLAN 之间的绑定

网络协议	VLAN ID
IP	VLAN 2
IPX	VLAN 3

4. 基于网络地址划分的 VLAN

基于网络地址划分的 VLAN 中的基本成员是终端，根据终端使用的网络地址来确定终端所属的 VLAN，为了确定某个终端所属的 VLAN，必须建立如表 2.4 所示的网络地址与 VLAN 之间的绑定，交换机根据接收到的 IP 分组的源 IP 地址和网络地址与 VLAN 之间的绑定来确定发送 IP 分组的终端所属的 VLAN。由于目前终端的 IP 地址通过 DHCP 自动获得，而且终端自动获得的 IP 地址往往取决于终端所属的 VLAN，因此，这种 VLAN 划分方式目前也很少使用。

表 2.4 网络地址和 VLAN 之间的绑定

网络地址	VLAN ID
192.1.1.0/24	VLAN 2
192.1.2.0/24	VLAN 3
192.1.3.0/24	VLAN 4
192.1.4.0/24	VLAN 5
192.1.5.0/24	VLAN 6
192.1.6.0/24	VLAN 7

4 种 VLAN 划分方式中，基于端口划分 VLAN 方式是最常用的 VLAN 划分方式，所有交换机都支持这种 VLAN 划分方式，基于 MAC 地址划分 VLAN 方式是比较高级的 VLAN 划分方式，各个厂家有着各自的基于 MAC 地址划分 VLAN 的技术，其他两种 VLAN 划分方式因为目前很少使用，不再展开讨论。

2.3 基于端口划分 VLAN

2.3.1 单交换机 VLAN 划分过程

基于端口划分 VLAN，首先需要在交换机中创建 VLAN，然后，将交换机端口分配给某个 VLAN，一般情况下，每一个接入端口只能分配给单个 VLAN。

图 2.6(a)是一个拥有 9 个端口的交换机，初始状态下，整个以太网交换机就是一个广播域，连接任何端口的终端所发送的广播帧(以目的 MAC 地址为广播地址的 MAC 帧)将从以太网交换机的所有其他端口发送出去，任何一个连接在该以太网交换机端口的终端都将收到该广播帧，那么如何才能分割广播域，使得广播帧只在少数几个端口内广播？比如说连接端口 1 的终端所发送的广播帧，只从以太网交换机的端口 3 和端口 5 发送出去，其他端口并不转发该广播帧。以太网交换机的 VLAN 功能可以使得以太网交换机能够用任意端口组合来构成一个广播域。在图 2.6(b)中，以太网交换机的端口 1、端口 3 和端口 5 构成一个广播域，以太网交换机的端口 2、端口 4 和端口 7 构成另一个广播域，而以太网交换机的剩余端口(端口 6、端口 8 和端口 9)又构成一个广播域，每一个广播域可以想象成一个用网桥连接的以太网。这样，将 9 个端口分割成 3 个广播域后的以太网交换机，逻辑上等同于在以太网交换机内设置了三个独立的网桥，这三个网桥分别连接属于三个不同广播域的端口，如图 2.6(c)所示。以太网交换机每个广播域的端口配置是任意的，因此，以太网交换机内的网桥也只有逻辑意义。一旦为以太网交换机配置了一个广播域，该广播域就拥有单独的转发表，属于该广播域的某个端口接收到 MAC 帧后，首先判别该 MAC 帧的目的 MAC 地址是否是广播地址，若是，就将该 MAC 帧从属于该广播域的所有其他端口发送出去，否则就用该 MAC 帧的目的 MAC 地址去查找转发表，如果找到匹配的转发项，就将该 MAC 帧从转发项指定的转发端口发送出去，MAC 帧的输入端口和输出端口必须属于同一个广播域，如果在转发表中找不到匹配的转发项，和广播帧一样，从属于该广播域的所有其他端口发送出去。

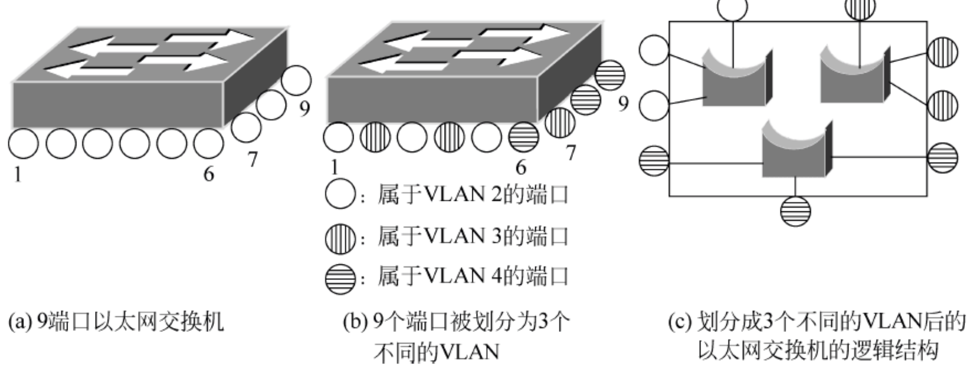


图 2.6 交换机划分 VLAN 过程

2.3.2　跨交换机 VLAN 划分过程

可以对以太网交换机任意配置广播域解决了动态分割广播域的问题，但分割的广播域仍然有着物理地域限制，真正不受物理地域限制的广播域划分是可以将一个由以太网交换机组成的大型交换式以太网的任意若干个端口组成一个广播域，如图 2.7 所示，这种划分广播域的技术称为跨以太网交换机划分 VLAN 技术。

如果两个位于不同以太网交换机的端口属于同一个 VLAN，则两个端口之间必须存在交换路径，假定端口 A 位于交换机 1，端口 B 位于交换机 2，为了建立端口 A 和端口 B 之间的交换路径，在交换机 1 中选择某个端口 C，它和端口 A 属于同一个 VLAN，且连接交换机 2 和端口 B 属于同一 VLAN 的某个端口 D。因此，交换机 1 中至少配置一个和端口 A 属于同一 VLAN 的端口 C，并使端口 C 连接交换机 2 的端口 D。交换机 2 也必须将端口 D 和端口 B 配置成属于同一 VLAN 的两个端口。保证属于同一 VLAN 的两个端口之间存在交换路径是跨交换机 VLAN 的配置规则。

如图 2.7 所示，为了实现终端 A 和终端 D 之间可以互相通信，终端 B 和终端 C 之间可以互相通信，终端 A、D 和终端 B、C 之间不能互相通信的目标，分别在交换机 1 和交换机 2 中将连接终端 A 和终端 D 的端口配置给 VLAN 2，连接终端 B 和终端 C 的端口配置给 VLAN 3。在每一个以太网交换机端口只能属于一个 VLAN 的情况下，必须在交换机 1 和交换机 2 选择两个端口，并将这两个端口分别配置给 VLAN 2 和 VLAN 3，用两条物理链路互连交换机 1 和交换机 2 中分别属于 VLAN 2 和 VLAN 3 的两对端口(注意：这里的每一条物理链路都是指全双工点对点信道)。但这样做也会带来一些问题，如果两个以太网交换机之间的物理距离很远，就需要配置用于实现以太网交换机之间互连的光端口，而且必须在以太网交换机之间铺设光缆，但每个以太网交换机的光端口数量和需要的光纤对数是不确定的，随着跨以太网交换机 VLAN 的数量变化而变化。如果需要对一个大型交换式以太网实现 VLAN 动态划分，用于以太网交换机之间互连的物理链路数更是不可预测，这仍将对网络的设计、实施带来困难。因此，实现跨以太网交换机 VLAN 划分必须解决的问题是通过以太网交换机之间单一的物理链路建立任何两个属于同一 VLAN 的端口之间的交换路径。

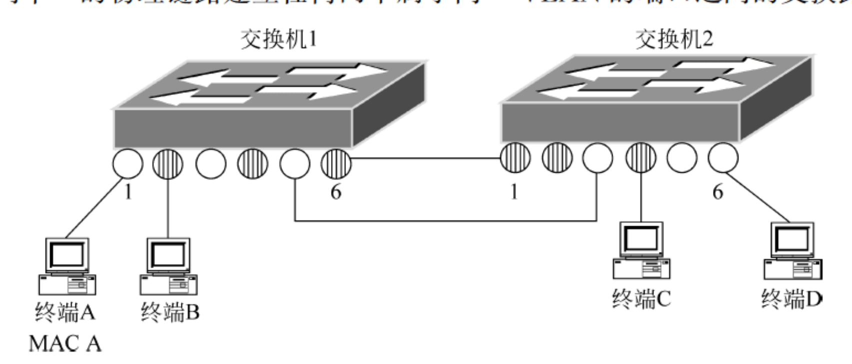


图 2.7　跨以太网交换机 VLAN 划分

2.3.3　802.1Q 与 VLAN 内数据传输

图 2.8 是用单一物理链路实现跨以太网交换机 VLAN 内终端之间通信的网络结构图，

图 2.9 是任意 VLAN 内两个终端之间完成通信的过程。

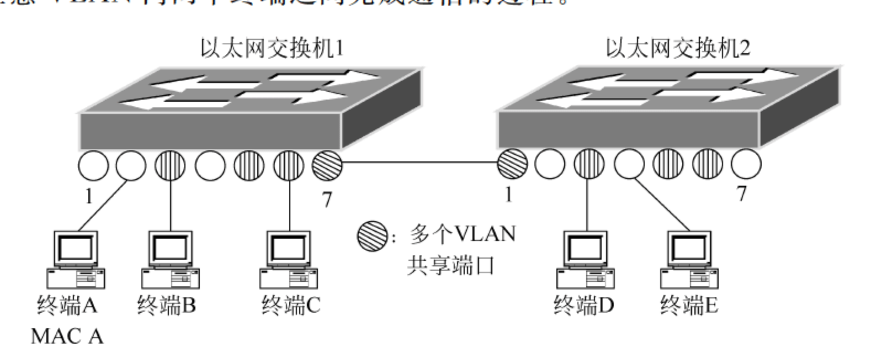


图 2.8 单一物理链路实现跨以太网交换机 VLAN 内终端之间通信

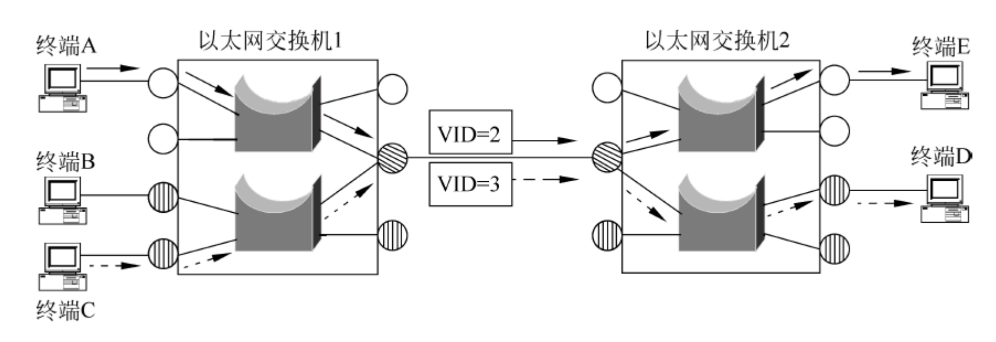


图 2.9 跨以太网交换机 VLAN 内终端之间实现通信的过程

针对图 2.8 所示的 VLAN 划分，交换机 1 端口 7 和交换机 2 端口 1 必须同时属于 VLAN 2 和 VLAN 3，这种同时属于多个 VLAN 的端口称为共享端口，连接这两个共享端口的物理链路必须成为任何一对属于同一 VLAN 的跨交换机端口之间的交换路径的一部分。但这样配置对实现终端 A→终端 E 之间通信有什么难度呢？当终端 A 通过端口 2 向以太网交换机 1 发送源 MAC 地址为 MAC A，目的 MAC 地址为 MAC E 的 MAC 帧时，以太网交换机 1 通过端口 2 接收到该 MAC 帧，确定在 VLAN 2 内转发该 MAC 帧，如果在和 VLAN 2 关联的转发表中找不到和 MAC E 匹配的转发项，以太网交换机 1 通过端口 1、端口 4 和端口 7 将该 MAC 帧转发出去。如果在和 VLAN 2 关联的转发表中找到和 MAC E 匹配的转发项，则只通过端口 7 转发该 MAC 帧。通过端口 7 转发出去的 MAC 帧通过以太网交换机 2 的端口 1 输入以太网交换机 2，由于以太网交换机 2 的端口 1 是共享端口，以太网交换机 2 无法根据接收该 MAC 帧的端口确定转发该 MAC 帧的 VLAN。

其实，以太网交换机 1 在将该 MAC 帧从其端口 7 转发出去时，是知道该 MAC 帧所属的 VLAN 的，而且以太网交换机 1 也知道端口 7 是共享端口，属于不同 VLAN 的 MAC 帧都有可能从该端口转发出去。为了让接收从该端口转发出去的 MAC 帧的设备能够确定每一个从其转发出去的 MAC 帧所属的 VLAN，以太网交换机 1 对所有从共享端口转发出去的 MAC 帧加上一个 VLAN 标识符字段(VID)，包含 VLAN 标识符字段的 MAC 帧的格式如图 2.10 所示。这种携带 VLAN 标识符字段的 MAC 帧结构称为 802.1Q 帧格式，802.1Q是 IEEE 802 委员会为实现跨以太网交换机的 VLAN 内的终端之间通信而制定的标准。图 2.10 中源 MAC 地址字段之后的 2 字节 8100H(本教材用 8100H 表示十六进制

8100)用于指明该 MAC 帧携带 VLAN 标识符,为与类型字段相区分,类型字段值中不允许出现 8100H。

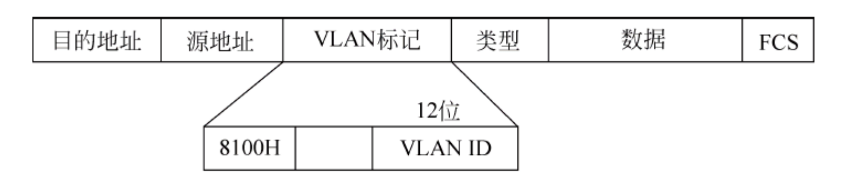


图 2.10 带 VLAN 标识符字段的 MAC 帧格式(802.1Q)

如图 2.9 所示,由于以太网交换机 1 通过共享端口(端口 7)转发 MAC 帧时,在 MAC 帧上加上了 VLAN 标识符(VID=2),当以太网交换机 2 通过共享端口(端口 1)接收到该 MAC 帧时,不是通过接收该 MAC 帧的端口,而是通过该 MAC 帧所携带的 VLAN 标识符(VID=2)确定用于转发该 MAC 帧的广播域(或 VLAN),并用该 MAC 帧携带的目的 MAC 地址查找和该广播域关联的转发表,如果在转发表中找到匹配的转发项,通过转发项给出的转发端口(端口 4)转发该 MAC 帧,否则,通过广播域内的所有其他端口(端口 2、端口 4 和端口 7)转发该 MAC 帧。

2.3.4 端口确定 MAC 帧所属 VLAN 规则

如果以太网交换机支持 802.1Q,单个端口可能属于多个 VLAN,为了确定从该端口输入的 MAC 帧所属的 VLAN,MAC 帧需要携带 VLAN 标识符,以太网交换机通过该 MAC 帧携带的 VLAN 标识符确定用于转发该 MAC 帧的 VLAN。为了标识从某个共享端口输出的 MAC 帧所属的 VLAN,需要给通过共享端口输出的 MAC 帧加上 VLAN 标识符。这种必须通过 MAC 帧携带的 VLAN 标识符确定用于转发该 MAC 帧的 VLAN 的端口被称为标记端口。而那些通过输入 MAC 帧的端口就能确定用于转发该 MAC 帧的 VLAN 的端口被称为非标记端口,接入端口通常是非标记端口。某个端口可以同时作为标记端口和非标记端口加入多个 VLAN,作为标记端口可以同时加入若干 VLAN,作为非标记端口只允许加入一个 VLAN。假定某个端口作为非标记端口加入了 VLAN 1,作为标记端口加入了 VLAN 2 和 VLAN 3。从该端口输入 MAC 帧时,首先判别该 MAC 帧是否携带 VLAN 标识符,如果携带 VLAN 标识符且 VLAN 标识符为 2 或 3,则确定 VLAN 2 或 VLAN 3 是用于转发该 MAC 帧的 VLAN。如果该 MAC 帧没有携带 VLAN 标识符,则确定 VLAN 1 是用于转发该 MAC 帧的 VLAN。其他情况下,丢弃该 MAC 帧。

2.3.5 VLAN 例题解析

【例 2.1】 交换机连接终端和集线器的方式及端口分配给各个 VLAN 的情况如图 2.11 所示,假定终端后面的字符表示终端的 MAC 地址,初始转发表为空表,回答以下问题。

① 终端 A 发送的目的 MAC 地址为 B 的 MAC 帧到达哪些终端?

② 终端 B 发送的目的 MAC 地址为 A 的 MAC 帧到达哪些终端?

③ 终端 E 发送的目的 MAC 地址为 B 的 MAC 帧到达哪些终端?

④ 终端 B 发送的目的 MAC 地址为 E 的 MAC 帧到达哪些终端?

⑤ 终端 B 发送的目的 MAC 地址为广播地址的 MAC 帧到达哪些终端？

⑥ 终端 F 发送的目的 MAC 地址为 E 的 MAC 帧到达哪些终端？

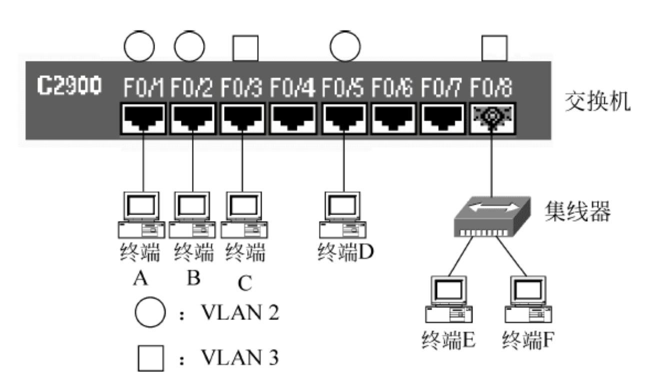


图 2.11 交换机连接终端和集线器的方式

【解析】 ① 由于初始转发表为空表，VLAN 2 对应的转发表中没有 MAC 地址 B 匹配的转发项，该 MAC 帧被交换机在 VLAN 2 内广播，到达 VLAN 2 内除终端 A 以外的所有其他终端(终端 B 和终端 D)。VLAN 2 对应的转发表中增加 MAC 地址＝A，输出端口(或转发端口)＝F0/1 的转发项。

② 由于 VLAN 2 对应的转发表中存在与 MAC 地址 A 匹配的转发项，交换机只从端口 F0/1 输出该 MAC 帧，该 MAC 帧只到达终端 A。VLAN 2 对应的转发表中增加 MAC 地址＝B，输出端口(或转发端口)＝F0/2 的转发项。

③ 由于连接终端 E 的设备是集线器，因此，该 MAC 帧被集线器广播，到达交换机端口 F0/8 和终端 F。由于 VLAN 3 对应的转发表中没有与 MAC 地址 B 匹配的转发项，该 MAC 帧被交换机在 VLAN 3 内广播，到达 VLAN 3 内除连接在端口 F0/8 以外的所有其他终端(终端 C)。因此，该 MAC 帧到达终端 C 和终端 F。VLAN 3 对应的转发表中增加 MAC 地址＝E，输出端口(或转发端口)＝F0/8 的转发项。

④ 由于 VLAN 2 对应的转发表中没有 MAC 地址 E 匹配的转发项，该 MAC 帧被交换机在 VLAN 2 内广播，到达 VLAN 2 内除终端 B 以外的所有其他终端(终端 A 和终端 D)。

⑤ 由于 MAC 帧的目的地址是广播地址，该 MAC 帧被交换机在 VLAN 2 内广播，到达 VLAN 2 内除终端 B 以外的所有其他终端(终端 A 和终端 D)。

⑥ 由于连接终端 F 的设备是集线器，该 MAC 帧被集线器广播，到达交换机端口 F0/8 和终端 E。由于 VLAN 3 对应的转发表中存在与 MAC 地址 E 匹配的转发项，且输出端口 F0/8 是交换机接收该 MAC 帧的端口，交换机丢弃接收到的 MAC 帧。该 MAC 帧只到达终端 E。

【例 2.2】 假定网络结构如图 2.12 所示，终端 A、终端 D 和终端 E 属于一个 VLAN (VLAN 2)，终端 B、终端 C 和终端 F 属于另一个 VLAN(VLAN 3)。

① 如何进行 VLAN 配置？

② 给出终端 B→终端 C、终端 A→终端 D、终端 F→终端 B 的传输过程。

③ 能否实现终端 B→终端 D 的通信？为什么？

【解析】 ①配置 VLAN 的原则是所有属于同一 VLAN 的终端之间必须存在交换路径，如果某个端口只有单个 VLAN 内终端之间的交换路径经过，则将该端口配置为该

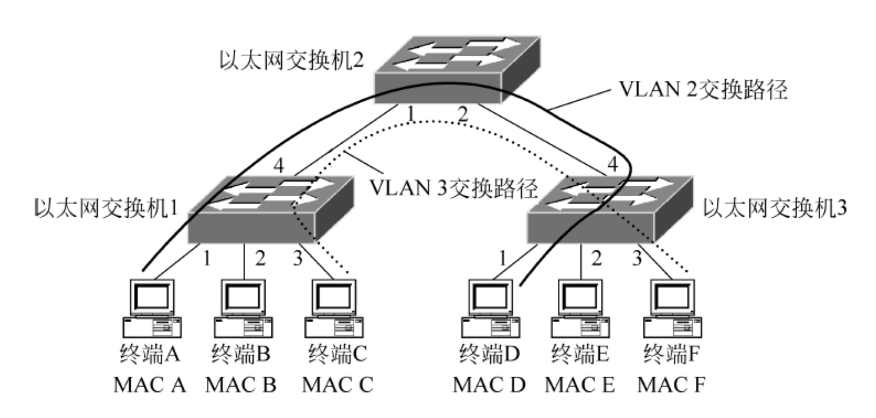


图 2.12 网络拓扑结构及 VLAN 划分

VLAN 的非标记端口；如果某个端口被多对属于不同 VLAN 的终端之间的交换路径经过，则将该端口配置为被这些 VLAN 共享的标记端口。根据图 2.12 给出的终端之间的交换路径，得出 VLAN 配置如图 2.13 所示。以太网交换机 1 配置两个 VLAN，分别命名为 VLAN 2 和 VLAN 3。VLAN 2 包括端口 1 和端口 4，其中端口 4 为标记端口，被两个 VLAN 共享。VLAN 3 包括端口 2、端口 3 和端口 4，端口 4 为标记端口。以太网交换机 2 配置两个 VLAN，端口 1 和端口 2 均被 VLAN 2 和 VLAN 3 所共享，因此，两个端口均是标记端口。以太网交换机 3 配置两个 VLAN，VLAN 2 包括端口 1、端口 2 和端口 4，VLAN 3 包括端口 3 和端口 4，端口 4 为标记端口，被 VLAN 2 和 VLAN 3 所共享。各个交换机 VLAN 与端口之间的关系如表 2.5 所示。

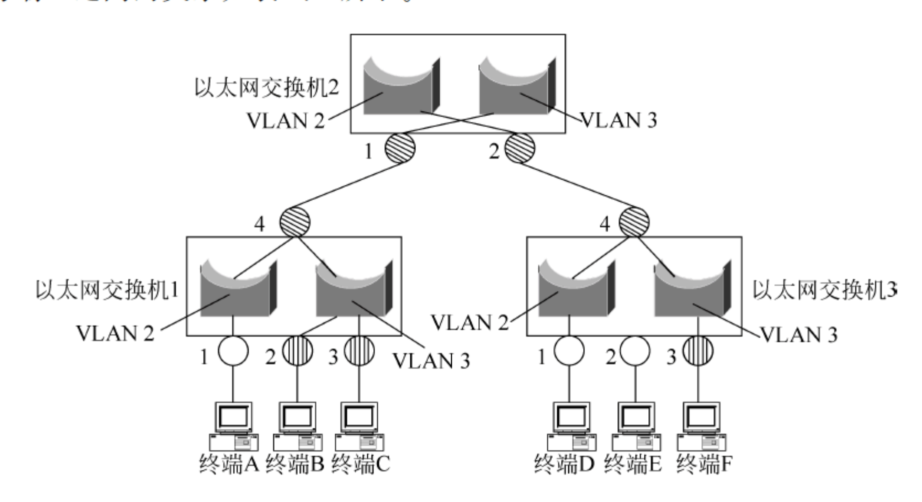


图 2.13 VLAN 配置

表 2.5 各个交换机 VLAN 与端口之间关系

交换机	VLAN 2		VLAN 3	
	非标记端口	标记端口	非标记端口	标记端口
以太网交换机 1	1.1	1.4	1.2,1.3	1.4
以太网交换机 2		2.1,2.2		2.1,2.2
以太网交换机 3	3.1,3.2	3.4	3.3	3.4

注：1.1 指以太网交换机 1 的端口 1。

② 在链路层传输 MAC 帧，必须事先知道源和目的终端的 MAC 地址，在本例中，假定终端 B 已经知道终端 C 的 MAC 地址为 MAC C，终端 B 构建一个以 MAC B 为源 MAC 地址，MAC C 为目的 MAC 地址的 MAC 帧，并将该 MAC 帧通过端口 2 发送给以太网交换机 1，以太网交换机 1 根据该 MAC 帧进入的端口（端口 2）确定该 MAC 帧在 VLAN 3 内传输，用目的 MAC 地址（MAC C）去查找和 VLAN 3 关联的转发表，由于没有找到匹配的转发项（一开始转发表为空表），在 VLAN 3 内广播该 MAC 帧，同时在以太网交换机 1 内和 VLAN 3 关联的转发表中添加 MAC 地址为 MAC B，转发端口为端口 2 这一转发项。对于以太网交换机 1 而言，属于 VLAN 3 的端口为端口 2、端口 3 和端口 4，因此，它通过除接收端口（端口 2）以外的所有其他端口（端口 3、端口 4）转发该 MAC 帧，由于端口 4 对于 VLAN 3 是 802.1Q 标记端口，从端口 4 转发出去的 MAC 帧需要携带 VLAN 标识符。因此，以太网交换机 1 在从端口 4 转发出去的 MAC 帧上加上 VLAN 3 的 VLAN 标识符（VID=3）。以太网交换机 2 通过端口 1 接收到该 MAC 帧，通过该 MAC 帧携带的 VLAN 标识符（VID=3）得知该 MAC 帧在 VLAN 3 内传输，同样用该 MAC 帧携带的目的 MAC 地址（MAC C）去查找和 VLAN 3 关联的转发表，也找不到匹配的转发项，以太网交换机 2 继续以广播方式传输该 MAC 帧，同时在和 VLAN 3 关联的转发表内添加该 MAC 帧源 MAC 地址（MAC B）对应的转发项，该 MAC 帧一直在图 2.14 所示的 VLAN 3 中广播，到达属于 VLAN 3 的所有终端，广播过程如图 2.14 所示。

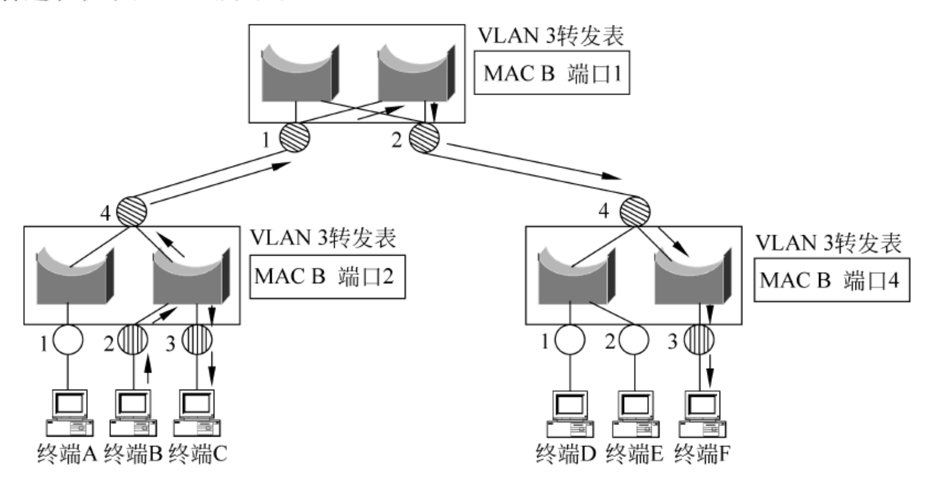


图 2.14 MAC 帧终端 B→终端 C 传输过程

终端 A→终端 D 的通信过程与终端 B→终端 C 的通信过程大致相同。由于以太网交换机 1、交换机 2、交换机 3 和 VLAN 1 关联的转发表中均没有与该 MAC 帧目的 MAC 地址（MAC D）匹配的转发项，因此，该 MAC 帧在 VLAN 2 内广播，广播过程如图 2.15 所示。

终端 F→终端 B 传输方式与前两次传输方式有所不同，由于以太网交换机 1、交换机 2、交换机 3 和 VLAN 3 关联的转发表中均有与该 MAC 帧目的 MAC 地址匹配的转发项，因此，以太网交换机 3 从端口 3 接收到该 MAC 帧后，只从端口 4 将该 MAC 帧转发出去，当然，转发出去的 MAC 帧携带 VLAN 3 的 VLAN 标识符（VID=3）。以太网交换机 2 通过查找和 VLAN 3 关联的转发表，将该 MAC 帧从端口 1 转发出去。以太网交换机 1 也通过查找和 VLAN 3 关联的转发表，将该 MAC 帧从端口 2 转发出去，由于在配置 VLAN 3 时指定端口 2 为非 802.1Q 标记端口，在将该 MAC 帧从端口 2 转发出去前，必须先删除该

MAC 帧上的 VLAN 标识符(VID=3),传输过程如图 2.16 所示。

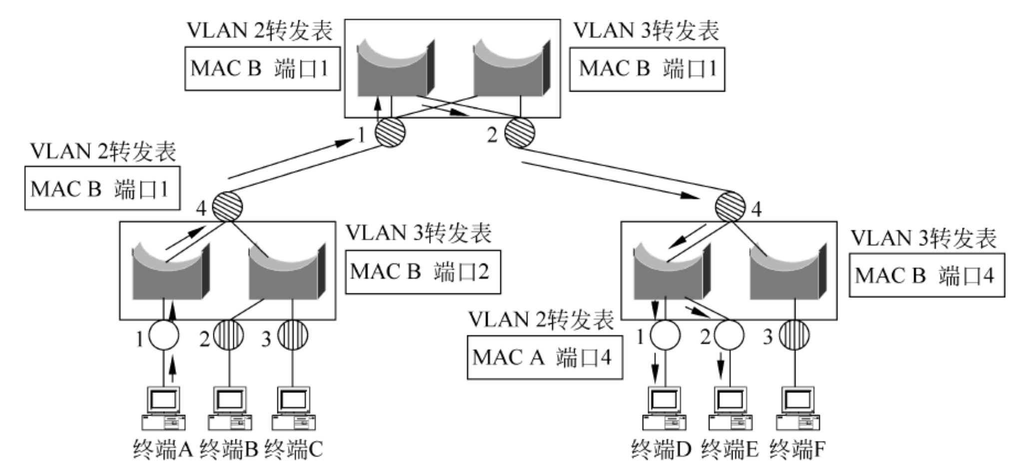


图 2.15 MAC 帧终端 A→终端 D 传输过程

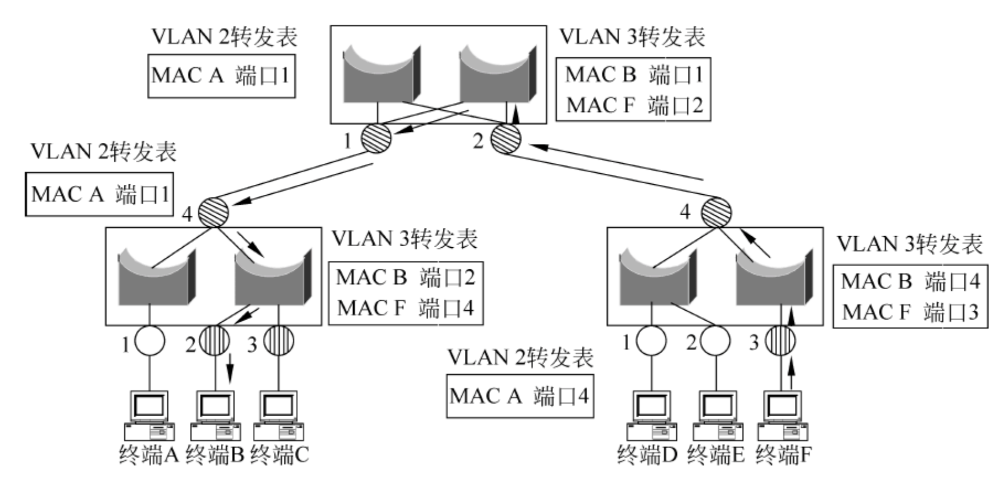


图 2.16 MAC 帧终端 F→终端 B 传输过程

③ 不能实现终端 B→终端 D 的通信。由于在和 VLAN 3 关联的转发表中找不到与 MAC D 匹配的转发项,以 MAC B 为源 MAC 地址、MAC D 为目的 MAC 地址的 MAC 帧只能以广播方式在 VLAN 3 内广播,但只能到达属于 VLAN 3 的所有终端,图 2.14 所示的是 MAC 帧终端 B→终端 C 传输过程。终端 D 属于 VLAN 2,该 MAC 帧到达不了终端 D。

【例 2.3】 VLAN 配置如图 2.17 所示,以太网交换机 1 的端口 1、端口 2、端口 4 和端口 7 属于 VLAN 2,端口 3、端口 5 和端口 6 属于 VLAN 3,所有端口均为非 802.1Q 标记端口。以太网交换机 2 的端口 2、端口 4 和端口 7 属于 VLAN 2,端口 1、端口 3、端口 5 和端口 6 属于 VLAN 3,所有端口均为非 802.1Q 标记端口。问:

① 终端 A 能否和终端 E 通信?为什么?

② 终端 B 能否和终端 D 通信?为什么?

③ 终端 A 能否和终端 D 通信?为什么?

④ 终端 B 能否和终端 E 通信?为什么?

【解析】 图 2.17 和图 2.8 的差别在于以太网交换机 1 的端口 7 和以太网交换机 2 的端口 1 的配置,在图 2.8 中,这两个端口均被 VLAN 2 和 VLAN 3 所共享,且都是 802.1Q

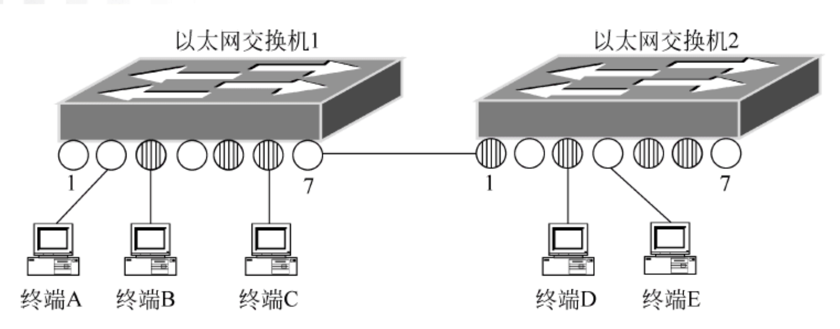


图 2.17 VLAN 配置图

标记端口，这种配置下，题中 4 问很容易回答，即使是跨以太网交换机，同一 VLAN 内的终端之间也可以互相通信，不同 VLAN 内的终端之间，即使连接在同一以太网交换机上，也不可以互相通信。但如果是如图 2.17 所示的 VLAN 配置方式，情况就不同了，对于①所要求的传输方式，由于终端 A 所连的端口 2 和端口 7 属于同一个 VLAN，终端 A 发送给终端 E 的 MAC 帧可以从端口 7 转发出去，进入以太网交换机 2 的端口 1。由于以太网交换机 1 的端口 7 是非 802.1Q 标记端口，进入以太网交换机 2 端口 1 的 MAC 帧没有携带任何 VLAN 标识信息，而以太网交换机 2 的端口 1 又属于 VLAN 3，该 MAC 帧被以太网交换机 2 在 VLAN 3 对应的广播域内进行转发，当然无法到达属于 VLAN 2 的端口 4，也就无法到达连接在端口 4 上的终端 E。对于②所要求的传输方式，由于终端 B 所连的端口 3 和端口 7 不属于同一个 VLAN，终端 B 发送的 MAC 帧无法从端口 7 转发出去，因而也无法进入以太网交换机 2 的端口 1，导致通信失败。对于③所要求的传输方式，由于终端 A 所连的端口 2 和端口 7 属于同一个 VLAN，终端 A 发送给终端 D 的 MAC 帧可以从端口 7 转发出去，进入以太网交换机 2 的端口 1。由于进入以太网交换机 2 的端口 1 的 MAC 帧没有携带任何 VLAN 标识信息，而以太网交换机 2 的端口 1 又属于 VLAN 3，该 MAC 帧被以太网交换机 2 在 VLAN 3 对应的广播域内进行转发，而终端 D 所连的端口 3 属于 VLAN 3，该 MAC 帧能够到达终端 D。对于④所要求的传输方式，由于终端 B 所连的端口 3 和端口 7 不属于同一个 VLAN，终端 B 发送的 MAC 帧无法从端口 7 转发出去，因而也无法进入以太网交换机 2 的端口 1，导致通信失败。造成上述情况的原因在于：如果输出端口是非 802.1Q 标记端口，VLAN 只有本地意义，如终端 A 发送的 MAC 帧，由于端口 2 属于 VLAN 2，被以太网交换机 1 在 VLAN 2 对应的广播域内进行转发，但一旦该 MAC 帧离开以太网交换机 1，就像终端 A 刚发送的 MAC 帧一样，由于没有携带任何 VLAN 标识信息，其他以太网交换机只能重新通过接收该 MAC 帧的端口来确定用于转发该 MAC 帧的 VLAN。

2.4 Cisco 基于 MAC 地址划分 VLAN 技术

到目前为止，讨论的 VLAN 划分方式都是基于端口的，在一个交换式以太网中，通过配置可以将任意端口组合定义成一个 VLAN，但这种端口组合是静态的，如果要改变某个 VLAN 的端口组合，必须重新对 VLAN 进行配置。或许存在这样的需求：允许一台笔记本计算机在校园漫游，不用重新配置该笔记本计算机的 IP 地址就可在校园各处上网。基于端口配置 VLAN 的方式对这种应用限制较大，如果该笔记本计算机配置了属于某个 IP 子网

的 IP 地址，那么，在不重新配置该笔记本计算机的 IP 地址的情况下，该笔记本计算机只能插入属于和该 IP 子网相关联的 VLAN 的端口，如假定该笔记本计算机的 IP 地址为 192.1.1.1，而和 IP 子网 192.1.1.0/24 关联的 VLAN 为 VLAN 7，如果该笔记本计算机要坚持使用 IP 地址 192.1.1.1，它只能插入属于 VLAN 7 的端口。如果这种需求很大，必须在每一楼层为所有 VLAN 预留一些平时不用的端口，当 VLAN 数目很大时，预留的端口数量就会很多。由于是静态配置且这些端口用于直接连接终端，每个端口只能固定对应一个 VLAN，因此，预留端口的利用率很低。

除了基于端口配置 VLAN 外，还有一种基于 MAC 地址动态配置 VLAN 的方法。假定某个交换机有 24 个端口，它可以用基于端口的配置方式将 22 个端口分配到指定 VLAN，但将余下的两个端口作为动态端口，这两个端口究竟属于哪一个 VLAN，由连接到端口上的终端的 MAC 地址确定，其过程如图 2.18 所示。

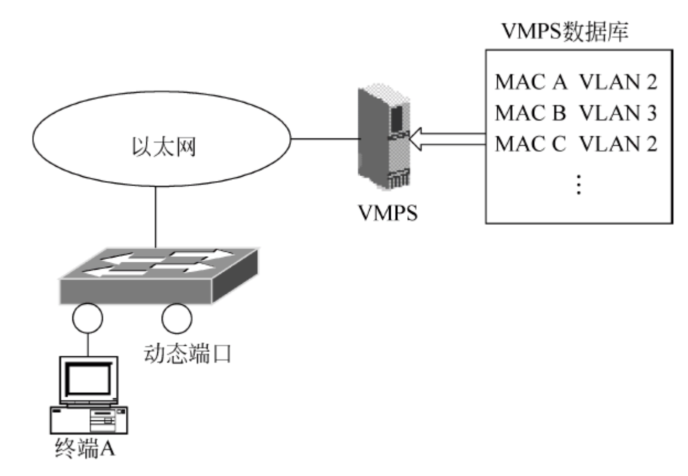


图 2.18　基于 MAC 地址划分 VLAN 过程

首先必须在 VLAN 成员策略服务器（VLAN Membership Policy Server，VMPS）建立 MAC 地址与 VLAN 之间的绑定，如图 2.18 中 VMPS 数据库所示。当某个动态端口接入终端，该终端就通过动态端口传输以该终端 MAC 地址（这里为 MAC A）为源 MAC 地址的 MAC 帧，以太网交换机接收到该 MAC 帧后，发现该端口是动态端口，且还没有将该端口配置给任何 VLAN，就向 VMPS 发送请求，请求中包含该 MAC 帧的源 MAC 地址，VMPS 用该 MAC 地址检索它的数据库，找到对应项，并确定和该 MAC 地址（MAC A）关联的 VLAN 是 VLAN 2。VMPS 向以太网交换机回送一个确认响应，并指出将该端口暂时配置给 VLAN 2。后续通过该端口接收到的 MAC 帧都在 VLAN 2 内进行转发。如果该端口一段时间内接收不到 MAC 帧，将重新回到初始状态，不再属于任何 VLAN，在再次能够转发 MAC 帧前，必须重新通过查询 VMPS 获得有关该端口所属 VLAN 的确认信息。

2.5　专用 VLAN

2.5.1　专用 VLAN 的作用

以太网作为接入网络的网络结构如图 2.19 所示。图 2.19 所示网络结构存在如何选择

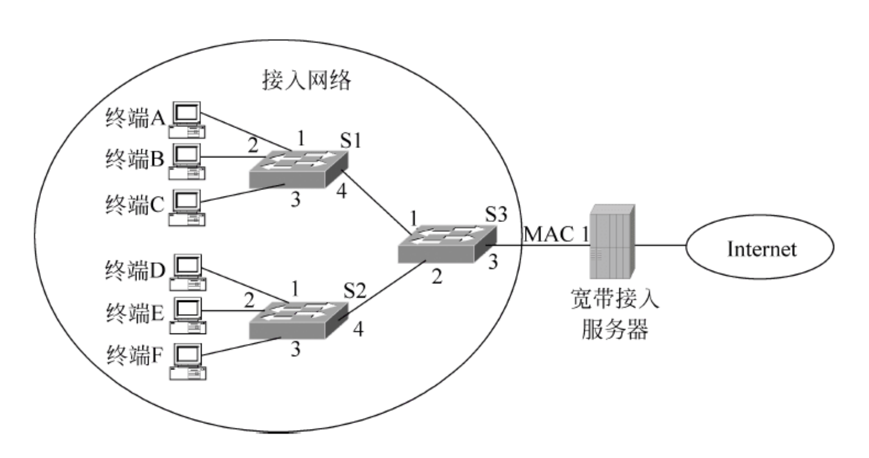


图 2.19 以太网接入网络

VLAN 划分单元的问题，如果将整个接入网络作为单个 VLAN，可以有效减少 VLAN 数量，同步减少宽带接入服务器中的路由项数量，只需为接入网络分配单个网络地址，所有用户终端有着相同的默认网关地址，但允许用户终端之间相互通信，每一个用户终端能够接收到以广播传输方式在接入网络中传输的 MAC 帧。如果将每一个连接用户终端的端口分配到不同的 VLAN，交换机之间互连的端口配置成共享端口，能够保证每一个用户终端只能与宽带接入服务器通信，但需要创建与用户终端数量相等的 VLAN。由于每一个 VLAN 对应着不同的网络，需要为每一个 VLAN 分配不同的网络地址，为每一个用户终端分配不同的默认网关地址，宽带接入服务器将同步增加路由项数量。能否有这样一种 VLAN 划分机制，在接入网络内部，每一个 VLAN 只包含连接用户终端的交换机端口（接入端口）和交换机连接其他交换机的端口（共享端口），保证每一个用户终端只能与宽带接入服务器通信？对于宽带接入服务器，这些 VLAN 又组合成一个单一 VLAN，宽带接入服务器将接入网络作为单个 VLAN 分配网络地址，建立路由项。这种 VLAN 划分技术就是专用 VLAN。

2.5.2 Cisco 专用 VLAN 工作原理

目前并不存在标准的专用 VLAN 协议，各个厂家实现专用 VLAN 的机制各不相同，这里以 Cisco 实现专用 VLAN 的技术为例讨论专用 VLAN 的工作原理。其他厂家的专用 VLAN 实现技术大致与此相似。

Cisco 专用 VLAN 实现技术采用二层 VLAN 机制，整个接入网络作为一个 VLAN，该 VLAN 作为主 VLAN，同时允许将主 VLAN 再次划分为多个次 VLAN。存在两种次 VLAN 类型，一是孤立 VLAN，二是团体 VLAN。每一个主 VLAN 只允许再次划分出一个孤立 VLAN，孤立 VLAN 中的接入端口称为孤立端口，属于同一孤立 VLAN 的孤立端口之间禁止相互通信。每一个主 VLAN 允许再次划分出多个团体 VLAN，团体 VLAN 中的接入端口称为团体端口，允许属于相同团体 VLAN 的团体端口之间相互通信。可以将主 VLAN 中的某个端口设置成混杂端口，允许孤立 VLAN 中的孤立端口与混杂端口相互通信，也允许不同团体 VLAN 中的团体端口与混杂端口相互通信。表 2.6 给出孤立端口、团体端口、混杂端口和共享端口之间的通信情况。

表 2.6 MAC 帧输入/输出端口之间的转发情况

输入端口	输出端口			
	孤立端口	团体端口	混杂端口	共享端口
孤立端口	禁止	禁止	允许(如果混杂端口是共享端口,携带主VLAN 对应的 VLAN ID)	允许(携带孤立VLAN 对应的VLAN ID)
团体端口	禁止	允许(属于和输入端口相同的团体 VLAN)	允许(如果混杂端口是共享端口,携带主VLAN 对应的 VLAN ID)	允许(携带团体VLAN 对应的VLAN ID)
共享端口	禁止	允许	允许(如果混杂端口是共享端口,携带主VLAN 对应的 VLAN ID)	允许(携带的 VLAN ID 不变)
混杂端口	允许(在主 VLAN 内转发,对于主VLAN,不存在孤立端口和团体端口)		允许(如果混杂端口是共享端口,携带主 VLAN 对应的 VLAN ID)	允许(携带主 VLAN 对应的 VLAN ID)

对于图 2.19 所示的接入网络,要求:①终端 B、终端 C、终端 D 和终端 E 只能与宽带接入服务器相互通信;②允许终端 A 和终端 F 相互通信,并与宽带接入服务器相互通信;③宽带接入服务器将整个接入网络作为单个 VLAN。下面以实现上述功能为例,讨论 Cisco 专用 VLAN 的工作原理。

1. 创建主次 VLAN

根据图 2.20 所示创建主次 VLAN。创建 VLAN ID 为 VLAN 100 的 VLAN,将其作为主 VLAN,创建 VLAN ID 为 VLAN 200 的 VLAN,将其作为孤立 VLAN,创建 VLAN ID 为 VLAN 300 的 VLAN,将其作为团体 VLAN。由于次 VLAN 是再次划分主 VLAN 的结果,因此,一是必须绑定主次 VLAN,二是所有属于次 VLAN 的端口均属于主 VLAN。

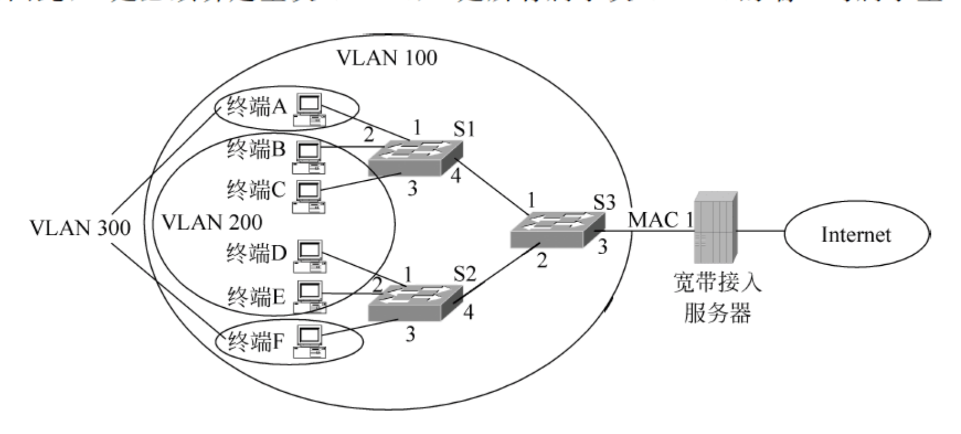


图 2.20 主 VLAN 和次 VLAN

2. 为孤立 VLAN 配置端口

将交换机 S1 端口 2 和端口 3、交换机 S2 端口 1 和端口 2 作为接入端口分配给 VLAN 200(次 VLAN 且孤立 VLAN),这些端口成为孤立 VLAN 中的孤立端口,将交换机 S1 端口 4、交换机 S2 端口 4、交换机 S3 端口 1 和端口 2 作为共享端口分配给 VLAN 200,这些端口成为孤立 VLAN 共享的共享端口。孤立端口输入的 MAC 帧禁止从其他孤立端口转发出去,但允许从孤立 VLAN 共享的共享端口转发出去,转发出去的 MAC 帧携带孤立 VLAN 对应的 VLAN ID。从孤立 VLAN 共享的共享端口接收到的 MAC 帧,如果携带孤立 VLAN 对应的 VLAN ID,只能从其他孤立 VLAN 共享的共享端口转发出去,禁止从其他孤立端口转发出去。

3. 为团体 VLAN 配置端口

将交换机 S1 端口 1、交换机 S2 端口 3 作为接入端口分配给 VLAN 300(次 VLAN 且团体 VLAN),这些端口成为团体 VLAN 中的团体端口,将交换机 S1 端口 4、交换机 S2 端口 4、交换机 S3 端口 1 和端口 2 作为共享端口分配给 VLAN 300,这些端口成为团体 VLAN 共享的共享端口。团体端口输入的 MAC 帧允许从属于同一团体 VLAN 的其他团体端口和该团体 VLAN 共享的共享端口转发出去,从团体 VLAN 共享的共享端口转发出去的 MAC 帧携带该团体 VLAN 对应的 VLAN ID。从团体 VLAN 共享的共享端口接收到的 MAC 帧,如果携带团体 VLAN 对应的 VLAN ID,允许从 VLAN ID 指定的团体 VLAN 共享的共享端口和属于 VLAN ID 指定的团体 VLAN 的其他团体端口转发出去。

4. 配置混杂端口

将交换机 S3 端口 3 作为主 VLAN 接入端口,并定义为混杂端口。从混杂端口接收到的 MAC 帧,如果从次 VLAN 共享的共享端口转发出去,携带的 VLAN ID 为主 VLAN 对应的 VLAN ID。混杂端口之间,混杂端口和孤立端口、团体端口、共享端口之间均能相互通信。从混杂端口转发出去的 MAC 帧完全等同于从主 VLAN 接入端口转发出去的 MAC 帧。需要强调的是,如果混杂端口是一个共享端口,无论是通过孤立端口,还是团体端口输入的 MAC 帧,通过混杂端口输出时,都携带主 VLAN 对应的 VLAN ID。

5. MAC 帧所属 VLAN

所有从孤立端口接收到的 MAC 帧,属于孤立 VLAN,在孤立 VLAN 内转发。所有从团体端口接收到的 MAC 帧,属于该团体端口所属的团体 VLAN,在该团体 VLAN 内转发。所有从共享端口接收到的 MAC 帧,根据 MAC 帧携带的 VLAN ID 确定该 MAC 帧所属的 VLAN,在 VLAN ID 指定的 VLAN 内转发。从混杂端口接收到的 MAC 帧,属于主 VLAN,在主 VLAN 内转发。主 VLAN 包含次 VLAN 包含的所有端口。

6. MAC 帧传输过程

假定终端 A～终端 F 对应的 MAC 地址为 MAC A～MAC F,交换机 S1～交换机 S3 已经建立如表 2.7～表 2.9 所示的各个 VLAN 对应的转发表,讨论下述终端之间、终端和宽带接入服务器之间 MAC 帧传输过程。

表 2.7　交换机 S1 中各个 VLAN 对应的转发表

VLAN ID	MAC 地址	转发端口	端口类型
VLAN 100	MAC A	端口 1	接入端口
VLAN 100	MAC B	端口 2	接入端口
VLAN 100	MAC C	端口 3	接入端口
VLAN 100	MAC D	端口 4	共享端口
VLAN 100	MAC E	端口 4	共享端口
VLAN 100	MAC F	端口 4	共享端口
VLAN 100	MAC 1	端口 4	共享端口
VLAN 200	MAC B	端口 2	孤立端口
VLAN 200	MAC C	端口 3	孤立端口
VLAN 200	MAC D	端口 4	共享端口
VLAN 200	MAC E	端口 4	共享端口
VLAN 300	MAC A	端口 1	团体端口
VLAN 300	MAC F	端口 4	共享端口

表 2.8　交换机 S2 中各个 VLAN 对应的转发表

VLAN ID	MAC 地址	转发端口	端口类型
VLAN 100	MAC A	端口 4	共享端口
VLAN 100	MAC B	端口 4	共享端口
VLAN 100	MAC C	端口 4	共享端口
VLAN 100	MAC D	端口 1	接入端口
VLAN 100	MAC E	端口 2	接入端口
VLAN 100	MAC F	端口 3	接入端口
VLAN 100	MAC 1	端口 4	共享端口
VLAN 200	MAC B	端口 4	共享端口
VLAN 200	MAC C	端口 4	共享端口
VLAN 200	MAC D	端口 1	孤立端口
VLAN 200	MAC E	端口 2	孤立端口
VLAN 300	MAC A	端口 4	共享端口
VLAN 300	MAC F	端口 3	团体端口

表 2.9　交换机 S3 中各个 VLAN 对应的转发表

VLAN ID	MAC 地址	转发端口	端口类型
VLAN 100	MAC A	端口 1	共享端口
VLAN 100	MAC B	端口 1	共享端口
VLAN 100	MAC C	端口 1	共享端口
VLAN 100	MAC D	端口 2	共享端口
VLAN 100	MAC E	端口 2	共享端口
VLAN 100	MAC F	端口 2	共享端口
VLAN 100	MAC 1	端口 3	混杂端口
VLAN 200	MAC B	端口 1	共享端口
VLAN 200	MAC C	端口 1	共享端口
VLAN 200	MAC D	端口 2	共享端口

续表

VLAN ID	MAC 地址	转发端口	端口类型
VLAN 200	MAC E	端口 2	共享端口
VLAN 300	MAC A	端口 1	共享端口
VLAN 300	MAC F	端口 2	共享端口

(1) 终端 B→终端 C。终端 B 发送的 MAC 帧通过端口 2 进入交换机 S1,由于端口 2 作为接入端口分配给 VLAN 200,且 VLAN 200 被定义为次 VLAN 和孤立 VLAN,因此,端口 2 为孤立端口,该 MAC 帧在 VLAN 200 内转发。交换机 S1 在表 2.7 中查找 VLAN ID 为 VLAN 200、MAC 地址为 MAC B 的转发项,匹配的转发项确定从端口 3 转发该 MAC 帧,由于端口 3 是孤立端口,交换机 S1 丢弃该 MAC 帧。

(2) 终端 B→终端 D。终端 B 发送的 MAC 帧通过端口 2 进入交换机 S1,由于端口 2 为孤立端口,该 MAC 帧在 VLAN 200 内转发。交换机 S1 在表 2.7 中查找 VLAN ID 为 VLAN 200、MAC 地址为 MAC D 的转发项,匹配的转发项确定从端口 4 转发该 MAC 帧,由于端口 4 是共享端口,交换机 S1 将该 MAC 帧从端口 4 转发出去,从端口 4 转发出去的 MAC 帧携带 VLAN ID——VLAN 200。

交换机 S3 从端口 1 接收到该 MAC 帧,由于端口 1 是共享端口,根据 MAC 帧携带的 VLAN ID 确定在 VLAN 200 内转发该 MAC 帧,交换机 S3 在表 2.9 中查找 VLAN ID 为 VLAN 200、MAC 地址为 MAC D 的转发项,匹配的转发项确定从端口 2 转发该 MAC 帧,由于端口 2 是共享端口,交换机 S3 将该 MAC 帧从端口 2 转发出去,从端口 2 转发出去的 MAC 帧携带 VLAN ID——VLAN 200。

交换机 S2 从端口 4 接收到该 MAC 帧,由于端口 4 是共享端口,根据 MAC 帧携带的 VLAN ID 确定在 VLAN 200 内转发该 MAC 帧,交换机 S2 在表 2.8 中查找 VLAN ID 为 VLAN 200、MAC 地址为 MAC D 的转发项,匹配的转发项确定从端口 1 转发该 MAC 帧,由于端口 1 是孤立端口,交换机 S2 丢弃该 MAC 帧。

(3) 终端 A→终端 F。终端 A 发送的 MAC 帧通过端口 1 进入交换机 S1,由于端口 1 作为接入端口分配给 VLAN 300,且 VLAN 300 被定义为次 VLAN 和团体 VLAN,因此,端口 1 为属于 VLAN ID 为 VLAN 300 的团体 VLAN 的团体端口,该 MAC 帧在 VLAN 300 内转发。交换机 S1 在表 2.7 中查找 VLAN ID 为 VLAN 300、MAC 地址为 MAC F 的转发项,匹配的转发项确定从端口 4 转发该 MAC 帧,由于端口 4 是共享端口,从端口 4 转发出去的 MAC 帧携带 VLAN ID——VLAN 300。

交换机 S3 从端口 1 接收到该 MAC 帧,由于端口 1 是共享端口,根据 MAC 帧携带的 VLAN ID 确定在 VLAN 300 内转发该 MAC 帧,交换机 S3 在表 2.9 中查找 VLAN ID 为 VLAN 300、MAC 地址为 MAC F 的转发项,匹配的转发项确定从端口 2 转发该 MAC 帧,由于端口 2 是共享端口,从端口 2 转发出去的 MAC 帧携带 VLAN ID——VLAN 300。

交换机 S2 从端口 4 接收到该 MAC 帧,由于端口 4 是共享端口,根据 MAC 帧携带的 VLAN ID 确定在 VLAN 300 内转发该 MAC 帧,交换机 S2 在表 2.8 中查找 VLAN ID 为 VLAN 300、MAC 地址为 MAC F 的转发项,匹配的转发项确定从端口 3 转发该 MAC 帧,由于端口 3 为属于 VLAN ID 为 VLAN 300 的团体 VLAN 的团体端口,交换机 S2 删除该

MAC 帧携带的 VLAN ID,将 MAC 帧从端口 3 转发出去,该 MAC 帧到达终端 F。

(4) 终端 A→宽带服务器。终端 A 发送的 MAC 帧通过端口 1 进入交换机 S1,由于端口 1 为属于 VLAN ID 为 VLAN 300 的团体 VLAN 的团体端口,该 MAC 帧在 VLAN 300 内转发。交换机 S1 在表 2.7 中找不到 VLAN ID 为 VLAN 300、MAC 地址为 MAC 1 的转发项,通过属于同一 VLAN 的其他团体端口、被 VLAN 300 共享的共享端口和混杂端口广播该 MAC 帧。该 MAC 帧从端口 4 转发出去,从端口 4 转发出去的 MAC 帧携带 VLAN ID——VLAN 300。

交换机 S3 从端口 1 接收到该 MAC 帧,由于端口 1 是共享端口,根据 MAC 帧携带的 VLAN ID 确定在 VLAN 300 内转发该 MAC 帧,交换机 S3 在表 2.9 中找不到 VLAN ID 为 VLAN 300、MAC 地址为 MAC 1 的转发项,通过属于同一 VLAN 的其他团体端口、被 VLAN 300 共享的共享端口和混杂端口广播该 MAC 帧。由于交换机 S3 的端口 3 既是混杂端口,又是主 VLAN 的接入端口,交换机 S3 删除该 MAC 帧携带的 VLAN ID,将该 MAC 帧从端口 3 转发出去,从端口 3 转发出去的 MAC 帧到达宽带接入服务器。由于交换机 S3 的端口 2 是 VLAN 300 的共享端口,从端口 1 接收到的携带 VLAN ID 为 VLAN 300 的 MAC 帧也通过端口 2 转发出去。从端口 2 转发出去的 MAC 帧到达交换机 S2。

交换机 S2 从端口 4 接收到该 MAC 帧,由于端口 4 是共享端口,根据 MAC 帧携带的 VLAN ID 确定在 VLAN 300 内转发该 MAC 帧,交换机 S2 在表 2.8 中找不到 VLAN ID 为 VLAN 300、MAC 地址为 MAC 1 的转发项,通过属于同一 VLAN 的其他团体端口、被 VLAN 300 共享的共享端口广播该 MAC 帧。由于交换机 S2 的端口 3 是属于 VLAN ID 为 300 的团体 VLAN 的团体端口,交换机 S2 删除该 MAC 帧携带的 VLAN ID,将该 MAC 帧从端口 3 转发出去。由于终端 F 的 MAC 地址不是 MAC 1,终端 F 丢弃该 MAC 帧。

(5) 宽带服务器→终端 D。宽带接入服务器发送的 MAC 帧通过端口 3 进入交换机 S3,由于端口 3 既是混杂端口,又是主 VLAN 接入端口,该 MAC 帧确定在主 VLAN 内转发。交换机 S3 在表 2.9 中查找 VLAN ID 为 VLAN 100、MAC 地址为 MAC B 的转发项,匹配的转发项确定从端口 2 转发该 MAC 帧,由于端口 2 是共享端口,从端口 2 转发出去的 MAC 帧携带 VLAN ID——VLAN 100。

交换机 S2 从端口 4 接收到该 MAC 帧,由于端口 4 是共享端口,根据 MAC 帧携带的 VLAN ID 确定在 VLAN 100 内转发该 MAC 帧,交换机 S2 在表 2.8 中查找 VLAN ID 为 VLAN 100、MAC 地址为 MAC D 的转发项,匹配的转发项确定从端口 1 转发该 MAC 帧,由于端口 1 是接入端口,交换机 S2 删除该 MAC 帧携带的 VLAN ID,将 MAC 帧从端口 1 转发出去。该 MAC 帧到达终端 D。

2.6 VLAN 属性注册协议

2.6.1 GVRP 作用

如果要实现图 2.21 所示的 VLAN 配置,需要在交换机 S1 创建 VLAN 2 和VLAN 3,将端口 1 作为接入端口分配给 VLAN 2,将端口 2 作为接入端口分配给 VLAN 3,将端口 3 作为被 VLAN 2 和 VLAN 3 共享的共享端口。交换机 S2 和交换机 S3 也需进行相似的配置过程,但交换机 S2 的端口 1 和端口 2 均作为被 VLAN 2 和 VLAN 3 共享的共享端口。

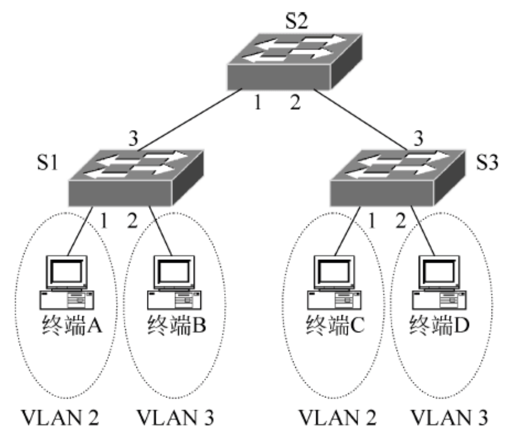


图 2.21 VLAN 配置过程

如果三台交换机的地域分布非常广泛，在三台交换机上手工创建 VLAN，并使得在三台交换机上手工创建的 VLAN 保持一致是困难的。能否有这样一种机制，只需在一台交换机上创建 VLAN，这些 VLAN 的属性能够自动分发到交换式以太网中的所有其他交换机，并在这些交换机上自动创建具有相同属性的 VLAN？VLAN 属性注册协议（GARP VLAN Registration Protocol，GVRP）就是这样一种机制。

2.6.2 GARP

1. GARP 简介

GVRP 是通用属性注册协议（Generic Attribute Registration Protocol，GARP）的一种应用。GARP 的作用是向其他参与者的应用实体提供注册属性和注销属性。图 2.22 中，交换机 S1、交换机 S2 和交换机 S3 都是参与者，每一个交换机端口都是 GARP 的应用实体，注册的属性随 GARP 应用的不同而不同，对于 GVRP，注册的属性是 VLAN 相关属性。

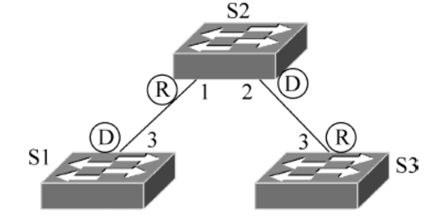


图 2.22 GARP 注册属性过程

GARP 通过声明消息和撤销声明消息来完成属性的注册和注销，图 2.22 中 S1 端口 3（声明端口）向 S2 端口 1 发送声明消息，S2 端口 1（注册端口）接收到声明消息后，在端口 1 中注册声明消息指定的属性。如果 S1 端口 3 向 S2 端口 1 发送撤销声明消息，S2 端口 1 接收到撤销声明消息后，在端口 1 中注销撤销声明消息指定的属性。注册和注销属性操作对于不同的 GARP 应用是不同的，对于 GVRP，注册属性是指将端口加入到指定 VLAN，注销属性是指将端口退出指定 VLAN。

2. GARP 消息类型

GARP 主要有三类消息，分别为 Join 消息、Leave 消息和 LeaveAll 消息，Join 消息是声明消息，Leave 消息和 LeaveAll 消息是撤销声明消息。

1）Join 消息

当一个 GARP 应用实体希望其他设备注册自己的属性信息时，对外发送 Join 消息。通

常在接收到其他实体发送的 Join 消息或本设备静态配置了某些需要其他 GARP 应用实体注册的属性时，向外发送 Join 消息。

Join 消息分为 JoinEmpty 和 JoinIn 两种，如果发送 Join 消息的应用实体本身没有注册该属性，发送 JoinEmpty 消息，如果发送 Join 消息的应用实体本身已经注册该属性，发送 JoinIn 消息。

2）Leave 消息

当一个 GARP 应用实体希望其他设备注销自己的属性信息时，对外发送 Leave 消息。通常在接收到其他实体发送的 Leave 消息或静态注销了某些需要其他 GARP 应用实体注销的属性时，向外发送 Leave 消息。

Leave 消息分为 LeaveEmpty 和 LeaveIn 两种，如果发送 Leave 消息的应用实体本身没有注册该属性，发送 LeaveEmpty 消息，如果发送 Leave 消息的应用实体本身已经注册该属性，发送 LeaveIn 消息。

3）LeaveAll 消息

每个设备启动后，将同时启动 LeaveAll 定时器，当该定时器超时（或溢出）后，所有应用实体将对外发送 LeaveAll 消息，LeaveAll 消息注销所有属性。LeaveAll 消息的作用是周期性地清除网络中的垃圾属性，一旦注销所有属性，必须通过接收其他应用实体发送的 Join 消息重新注册属性，一些没有重新注册的属性，就是通过 LeaveAll 消息注销的垃圾属性，例如，某个属性已经被某个设备删除，但由于该设备突然断电，并没有发送 Leave 消息来通知其他应用实体注销该属性，导致该属性被这些应用实体长期错误注册，LeaveAll 消息可以注销这些属性。

3. 定时器

GARP 协议中用到了 4 个定时器，下面分别介绍一下它们的作用。

1）Join 定时器

Join 定时器用来控制 Join 消息（包括 JoinIn 消息和 JoinEmpty 消息）的发送过程。为了保证 Join 消息能够可靠地传输到其他应用实体，发送第一个 Join 消息后将启动 Join 定时器，如果在 Join 定时器溢出前接收到 JoinIn 消息，不需重发 Join 消息，如果在 Join 定时器溢出前没有接收到 JoinIn 消息，重发 Join 消息。

每个端口维护独立的 Join 定时器。

2）Hold 定时器

Hold 定时器用来减少 Join 消息（包括 JoinIn 和 JoinEmpty）和 Leave 消息（包括 LeaveIn 和 LeaveEmpty）的发送频率。当应用实体接收到其他应用实体发送的消息，或者设备配置或删除了某些需要通知其他应用实体的属性，应用实体不是立即向外发送消息，而是启动 Hold 定时器，设备对该时间段内接收到的消息进行合并，并将该时间段内配置或删除的属性尽可能插入合并后的消息中，在 Hold 定时器溢出后，发送合并后的消息，以此减少消息的发送频率。如果没有 Hold 定时器的话，每接收一个消息就发送一个消息，或者设备每完成一次配置或删除属性操作就发送一个消息，会大大增加网络中消息的数量。

每个端口维护独立的 Hold 定时器。Hold 定时器的值要小于等于 Join 定时器值的一半。

3）Leave 定时器

Leave 定时器用来控制属性注销。每个应用实体接收到 Leave 或 LeaveAll 消息后，启

动 Leave 定时器,如果直到 Leave 定时器溢出,都没有接收到该属性的 Join 消息,该属性才会被注销。因为某个属性可能有多个属性源,接收到其中一个属性源发送的 Leave 消息,只能表示该属性源已经删除该属性,并不代表所有的属性源都已经删除该属性,因此不能立刻注销该属性,而是要等待其他属性源可能发送的 Join 消息。如果某个属性源存在该属性,接收到 Leave 或 LeaveAll 消息后,将发送 Join 消息,因此,存在接收到 Leave 消息之后,接收到 Join 消息的可能,一旦发生这种情况,表示该属性仍然需要保留,不能注销。因此,只有在接收到 Leave 或 LeaveAll 消息后,超过两倍 Join 定时器时间后仍没有收到该属性的 Join 消息时,才能认为网络中该属性的所有属性源均已删除该属性,才允许注销该属性。这就要求 Leave 定时器的值大于两倍 Join 定时器的值。

每个端口维护独立的 Leave 定时器。

4) LeaveAll 定时器

每个设备启动后,将同时启动 LeaveAll 定时器,当该定时器溢出时,所有 GARP 应用实体将对外发送 LeaveAll 消息,随后再次启动 LeaveAll 定时器,开始新的一轮循环。接收到 LeaveAll 消息的实体将重新启动所有的定时器,包括 LeaveAll 定时器。任何设备只有在 LeaveAll 定时器溢出时,才向外发送 LeaveAll 消息,这样就避免了多个设备短时间内发送多个 LeaveAll 消息的情况。如果不同设备的 LeaveAll 定时器同时溢出,就会同时发送多个 LeaveAll 消息,增加不必要的 LeaveAll 消息数量,为了避免不同设备同时发生 LeaveAll 定时器溢出的情况,实际定时器值是在设置的 LeaveAll 定时器值与 1.5 倍设置的 LeaveAll 定时器值之间随机选择的一个值。

一次 LeaveAll 事件相当于全网所有属性的一次清零。由于 LeaveAll 影响范围很广,LeaveAll 定时器的值不能太小,至少应该大于 Leave 定时器的值。每个设备只在全局维护一个 LeaveAll 定时器。

2.6.3 GVRP 工作原理

1. 端口注册模式

交换机手工配置的 VLAN 为静态 VLAN,通过 GVRP 创建的 VLAN 为动态 VLAN,GVRP 能够在交换机中创建动态 VLAN,并将端口分配给动态创建的 VLAN。GVRP 有三种注册模式,不同的注册模式对静态 VLAN 和动态 VLAN 的处理方式也不同。

1) Normal 模式

允许动态 VLAN 在端口上进行注册,同时会发送静态 VLAN 和动态 VLAN 的声明消息。

2) Fixed 模式

不允许动态 VLAN 在端口上注册,只发送静态 VLAN 的声明消息。

3) Forbidden 模式

不允许动态 VLAN 在端口上进行注册,同时删除端口上除 VLAN 1(默认 VLAN)以外的所有其他 VLAN,只发送 VLAN 1 的声明消息。

2. GVRP 工作过程

下面以完成图 2.21 要求的 VLAN 配置为例,讨论 GVRP 的工作过程。

1）完成 GVRP 配置

在 GVRP 开始工作之前，需要在交换机 S1、交换机 S2 和交换机 S3 上使能 GVRP，同时，将交换机 S1 端口 3、交换机 S2 端口 1 和端口 2、交换机 S3 端口 3 配置为被所有 VLAN 共享的共享端口，在这些共享端口上使能 GVRP，并将这些共享端口的注册模式配置为 Normal 模式。

2）在交换机 S1 上手工创建静态 VLAN——VLAN 2

交换机 S1 手工创建静态 VLAN——VLAN 2 导致交换机 S1 使能 GVRP 的端口（端口 3）向外发送 Join 消息。由于端口 3 本身没有接收过针对 VLAN 2 的 Join 消息，没有注册 VLAN 2 属性，因此，发送的是 JoinEmpty 消息，封装 JoinEmpty 消息的 MAC 帧的目的地址是组地址 01-80-C2-00-00-21，图 2.23 中用 JE（VLAN 2）表示针对 VLAN 2 的 JoinEmpty 消息。各个交换机端口注册 VLAN 2 属性和转发 JoinEmpty 消息的过程如图 2.23 所示。交换机 S2 端口 1 接收到 JE（VLAN 2）后，创建动态 VLAN——VLAN 2，注册 VLAN 2 属性，并通过端口 2 向外发送 JE（VLAN 2）。交换机 S3 端口 3 接收到 JE（VLAN 2）后，创建动态 VLAN——VLAN 2，注册 VLAN 2 属性。此时交换机 S1、交换机 S2 和交换机 S3 均存在 VLAN 2，其中交换机 S1 中的 VLAN 2 是静态 VLAN，不能通过 Leave 和 LeaveAll 消息注销。交换机 S2 和交换机 S3 中的 VLAN 2 是动态 VLAN，可以通过 Leave 和 LeaveAll 消息注销。需要强调的是，交换机 S1 手工创建 VLAN 2 所引发的 VLAN 2 属性声明和注册过程，只是在交换机 S2 端口 2 和交换机 S3 端口 3 注册 VLAN 2 属性，即将端口加入 VLAN 2。交换机 S1 端口 3 和交换机 S2 端口 2 只是声明端口，并没有通过 GVRP 加入到 VLAN 2。但在初始配置时，已将交换机使能 GVRP 的端口配置成被所有 VLAN 共享的共享端口，因此，虽然 GVRP 并没有在交换机 S1 端口 3 和交换机 S2 端口 2 注册 VLAN 2 属性，但在交换机 S1、交换机 S2 和交换机 S3 中创建 VLAN 2 后，交换机 S1 和交换机 S3 之间仍然存在属于 VLAN 2 的交换路径。

交换机 S1 端口 3 针对 VLAN 2 发送两个 JE（VLAN 2）消息（或者发送一个 JE（VLAN 2）后，接收到一个针对 VLAN 2 的 JoinIn 消息），并用计数器记录这一发送状态，一旦接收到 LeaveAll 消息，清除端口 3 针对所有属性的 Join 消息发送状态，因此，在静态 VLAN——VLAN 2 依然存在的情况下，交换机 S1 端口 3 将再次启动针对 VLAN 2 的 Join 消息发送过程，即或者发送两个 JE（VLAN 2）消息，或者发送一个 JE（VLAN 2）后，接收到一个针对 VLAN 2 的 JoinIn 消息。

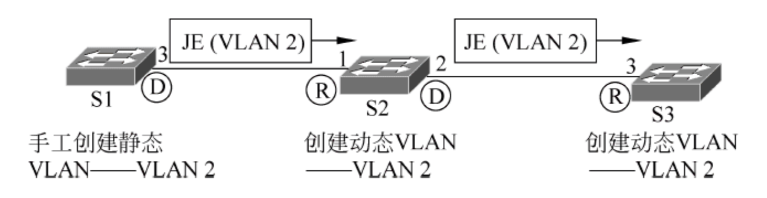


图 2.23　VLAN 2 属性声明和注册过程

3）在交换机 S3 上手工创建静态 VLAN——VLAN 3

交换机 S3 手工创建静态 VLAN——VLAN 3 后，进行图 2.24 所示的 VLAN 属性（VLAN 3）声明和注册过程。完成如图 2.24 所示的属性声明和注册过程后，交换机 S1、交换机 S2 和交换机 S3 均存在 VLAN 3，其中交换机 S3 中的 VLAN 3 是静态 VLAN，不能通

过 Leave 和 LeaveAll 消息注销。交换机 S2 和 S1 中的 VLAN 3 是动态 VLAN，可以通过 Leave 和 LeaveAll 消息注销。交换机 S3 端口 3 一旦接收到 LeaveAll 消息，同样将清除端口 3 针对所有属性的 Join 消息发送状态，因此，在静态 VLAN——VLAN 3 依然存在的情况下，交换机 S3 端口 3 将再次启动针对 VLAN 3 的 Join 消息发送过程。

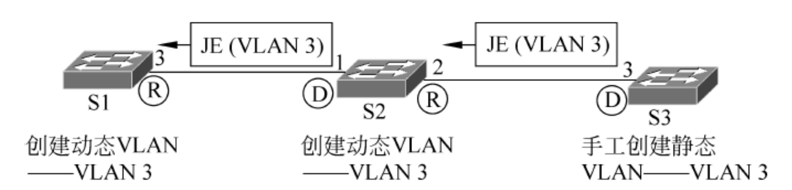


图 2.24 VLAN 3 属性声明和注册过程

4）交换机 S2 LeaveAll 定时器溢出

一旦交换机 S2 LeaveAll 定时器溢出，开始如图 2.25 所示的 LeaveAll 定时器溢出事件处理过程，交换机 S2 注销使能 GVRP 的端口（端口 1 和端口 2）已经注册的所有属性，并删除动态 VLAN——VLAN 2 和 VLAN 3，并通过使能 GVRP 的端口向外发送 LeaveAll 消息。交换机 S1 端口 3 接收到 LeaveAll 消息，注销 VLAN 3 属性，删除动态 VLAN——VLAN 3，清零端口 3 针对 VLAN 2 的 Join 消息发送状态，导致交换机 S1 再次发送针对 VLAN 2 的 JoinEmpty 消息，开始如图 2.23 所示的 VLAN 2 属性声明和注册过程。交换机 S3 端口 3 接收到 LeaveAll 消息，注销 VLAN 2 属性，删除动态 VLAN——VLAN 2，清零端口 3 针对 VLAN 3 的发送状态，导致交换机 S3 再次发送针对 VLAN 3 的 JoinEmpty 消息，开始如图 2.24 所示的 VLAN 3 属性声明和注册过程。完成如图 2.23 和图 2.24 所示的属性声明和注册过程后，交换机 S1、交换机 S2 和交换机 S3 中均创建 VLAN 2 和 VLAN 3，其中交换机 S1 中的 VLAN 2 和交换机 S3 中的 VLAN 3 是静态 VLAN。

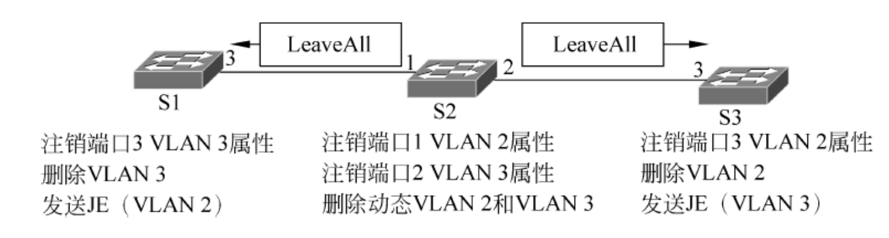


图 2.25 交换机 S2 LeaveAll 定时器溢出事件处理过程

5）在交换机 S3 上手工删除静态 VLAN——VLAN 3

一旦在交换机 S3 中手工删除静态 VLAN——VLAN 3，交换机 S3 将通过端口 3 发送针对 VLAN 3 的 LeaveEmpty 消息，图 2.26 中用 LE(VLAN 3)表示。交换机 S2 通过端口 2 接收到 LE(VLAN 3)消息后，在端口 2 中注销 VLAN 3 属性，由于 VLAN 3 属性只在端口 2 中注册，因此，交换机 S2 删除动态 VLAN——VLAN 3。完成这些操作后，交换机 S2 通过端口 1 向外发送 LE(VLAN 3)消息。交换机 S1 通过端口 3 接收到 LE(VLAN 3)消息后，在端口 3 中注销 VLAN 3 属性，由于 VLAN 3 属性只在端口 3 中注册，因此，交换机 S1 删除动态 VLAN——VLAN 3。完成如图 2.26 所示的 VLAN 3 属性注销过程后，交换机 S1、交换机 S2 和交换机 S3 中只存在 VLAN 2，其中交换机 S1 中的 VLAN 2 是手工创建的静态 VLAN。

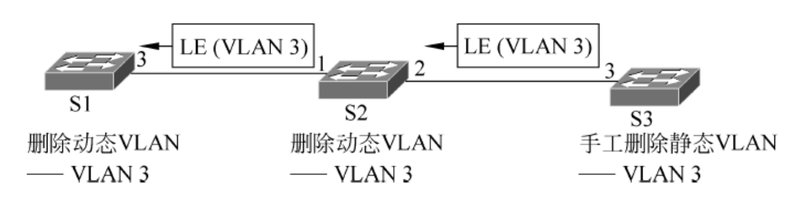


图 2.26 VLAN 3 属性注销过程

需要强调的是，如果某个端口注册了某个 VLAN 属性，但交换机不存在该 VLAN，必须创建该 VLAN，由于该 VLAN 由 GVRP 创建，是动态 VLAN。如果交换机中所有端口都注销了某个动态 VLAN 属性，交换机删除该动态 VLAN。

GVRP 使得在一台交换机上手工创建静态 VLAN 的操作通过属性注册被快速传播到其他交换机上，同样，一台交换机上手工删除某个静态 VLAN 的操作，也通过属性注销被快速传播到其他交换机上，通过在一台交换机上手工创建和删除 VLAN，使交换式以太网中的所有其他交换机自动创建和删除与该交换机一致的 VLAN，简化了交换式以太网的 VLAN 配置。

2.6.4 VTP

VLAN 主干协议（VLAN Trunking Protocol，VTP）是 Cisco 专用协议，其作用与 GVRP 相似。首先创建一个 VTP 域，在一台属于该 VTP 域的交换机上创建和删除 VLAN 的操作可以扩散到整个 VTP 域，即属于同一 VTP 域的交换机自动创建和删除与该交换机一致的 VLAN。这就保证，在整个 VTP 域中，只需对一台属于该 VTP 域的交换机进行 VLAN 配置，其配置结果可以扩散到属于同一 VTP 域的所有其他交换机。这将大大简化交换式以太网的 VLAN 配置。

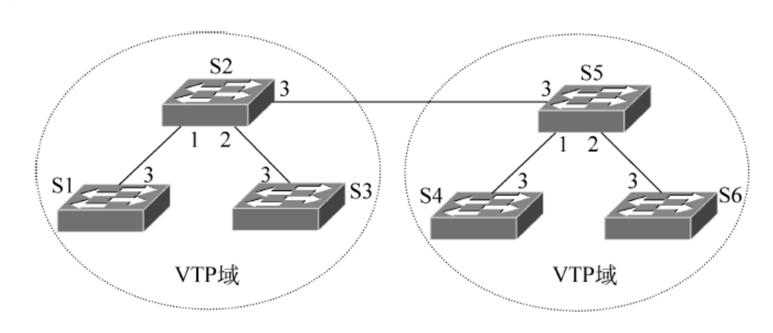


图 2.27 VTP 域

1. VTP 域

VTP 域由一组域名配置相同，通过共享端口互连的交换机组成。属于同一 VTP 域的交换机之间必须通过共享端口互连，因此，只有当图 2.27 中交换机 S1 端口 3、交换机 S2 端口 1 和端口 2、交换机 S3 端口 3 是共享端口时，才能保证交换机 S1、交换机 S2 和交换机 S3 构成一个 VTP 域。在实际配置过程中，为某台交换机配置的域名将自动传播给通过共享端口互连的一组交换机，如果图 2.27 中所有用于互连交换机的端口都是共享端口，在其中一台交换机上配置的域名将传播给图 2.27 中的所有交换机。为了生成图 2.27 所示的两个 VTP 域，必须在域边缘交换机（交换机 S2 或交换机 S5）手工配置不同的域名，如在交换机

S5 配置域名 ABC 后,在交换机 S2 配置域名 BCD。

需要强调的是,VTP 域只是确定了 VTP 消息的传播范围,与广播域分割无关,如果属于不同的 VTP 域的两个端口属于编号(VLAN ID)相同的 VLAN,且这两个端口之间存在属于该 VLAN 的交换路径,这两个端口属于同一个广播域,端口之间可以相互通信。

2. 交换机模式

VTP 将交换机模式分为服务器、客户和透明。

1) 服务器模式

服务器模式交换机允许创建和删除 VLAN,创建和删除 VLAN 的操作传播到属于同一 VTP 域的所有其他交换机,这些交换机自动创建和删除与该交换机一致的 VLAN。服务器模式交换机既可以手工创建和删除 VLAN,也可以根据相邻交换机发送给它的 VTP 消息自动创建和删除 VLAN,一旦交换机中 VLAN 发生改变,立即通过共享端口向相邻交换机发送 VTP 消息。在没有 VLAN 发生变化的情况下,服务器模式交换机定期通过共享端口向相邻交换机发送 VTP 消息。

2) 客户模式

客户模式交换机不允许手工创建和删除静态 VLAN,只有接收到 VTP 消息后,才可根据 VTP 消息创建和删除动态 VLAN。客户模式交换机完成动态 VLAN 创建和删除的操作后,立即通过共享端口向相邻交换机发送 VTP 消息,在没有 VLAN 发生变化的情况下,定期通过共享端口向相邻交换机发送 VTP 消息。客户模式交换机创建的动态 VLAN,断电后不会保留。

3) 透明模式

透明模式交换机只是转发 VTP 消息,不对 VTP 消息作任何处理。当它从一个共享端口接收到 VTP 消息后,它只是将 VTP 消息通过除接收该 VTP 消息以外的所有其他共享端口发送出去,VTP 消息本身对它是透明的。透明模式交换机允许手工创建和删除静态 VLAN,但手工创建和删除静态 VLAN 的操作只有本地意义,不对其他交换机产生影响。但如果透明模式交换机和其他模式交换机存在属于相同编号的 VLAN 的端口,这些端口属于同一个广播域。因此,交换式以太网中,所有属于相同编号的 VLAN 的端口属于同一个广播域,无论这些端口是分布在多个属于不同 VTP 域的交换机上,还是分布在多个有着不同模式的交换机上。

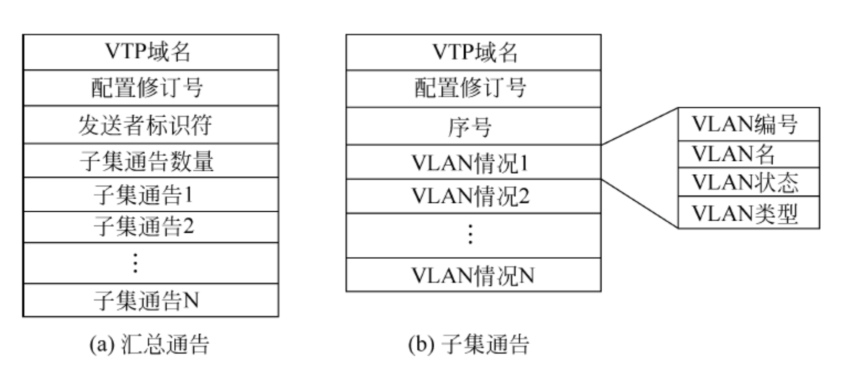
3. VTP 消息类型

VTP 定义了 4 种消息类型:汇总通告(Summary Advertisements)、子集通告(Subset Advertisements)、通告请求(Advertisement Requests)和 VTP Join 消息。前三种 VTP 消息用于实现同一 VTP 域中 VLAN 同步,最后一种消息用于 VTP 剪枝。这些消息在 VLAN 1 中传播,封装成 MAC 帧后的目的 MAC 地址为组播地址 01-00-0C-CC-CC-CC。

汇总通告和子集通告消息用于向相邻交换机通告 VLAN 情况,消息包含的主要信息如图 2.28 所示。

1) VTP 域名

每一个交换机只能属于一个 VTP 域,发送汇总通告和子集通告的交换机将自己的



(a) 汇总通告　　(b) 子集通告

图 2.28　汇总通告和子集通告包含的主要信息

VTP 域名作为汇总通告和子集通告的 VTP 域名。接收汇总通告和子集通告的交换机只有当自己的 VTP 域名与消息的 VTP 域名相同时，才对消息进行处理，否则，丢弃接收到的汇总通告和子集通告。

2）配置修订号

交换机的初始配置修订号为 1，只有当交换机 VLAN 情况发生变化时，才递增配置修订号，最大配置修订号的汇总通告和子集通告反映出交换机最新的 VLAN 情况。由于交换机只在两种情况下发送汇总通告和子集通告，一是交换机 VLAN 情况发生变化，二是超过两次发送汇总通告和子集通告的时间间隔定时器溢出。因此，即使在没有发生 VLAN 变化的情况下，交换机也定时发送汇总通告和子集通告，这些汇总通告和子集通告中的配置修订号维持不变。交换机对每一个向其发送汇总通告和子集通告的相邻交换机保留最新的配置修订号，当接收到某个相邻交换机发送的汇总通告和子集通告，并且消息中给出的配置修订号大于为该汇总通告和子集通告发送者保留的配置修订号，用消息中的配置修订号替换保留的配置修订号，继续处理该消息，否则，丢弃该消息。

3）发送者标识符

汇总通告中给出发送汇总通告的交换机的标识符，交换机通常把为其配置的管理 IP 地址作为其标识符。接收到汇总通告和子集通告的交换机通过消息中发送者标识符检索为该消息发送者保留的最新配置修订号。

4）VLAN 情况

给出交换机目前的 VLAN 配置情况，如存在哪些 VLAN，这些 VLAN 的类型（这里主要是以太网 VLAN）、编号（VLAN ID）、名字和状态（激活或删除）等。

交换机只有当 VLAN 情况发生变化时，才在汇总通告后面跟随子集通告，汇总通告后面允许跟随多个子集通告，子集通告数量字段值给出跟随的子集通告数量。通过子集通告给出本次 VLAN 变化的结果。对于周期性发送的汇总通告，并不需要跟随子集通告，子集通告数量字段值为 0。

交换机在下述情况下将发送通告请求。

- 交换机重新启动。
- 接收到的汇总通告中的配置修订号大于为该消息发送者保留的配置修订号，但该汇总通告没有跟随子集通告或者跟随的子集通告不完整。
- 交换机配置新的 VTP 域名。

通告请求可以要求发送所有子集通告(即被请求交换机目前所有 VLAN 的情况),也可以只要求发送指定序号的子集通告。

4. VTP 工作过程

1) 初始配置

如果需要通过 VTP 实现图 2.21 所示的 VLAN 配置,且要求交换机 S1 和交换机 S3 均能创建和删除 VLAN,对交换机 S1、交换机 S2 和交换机 S3 完成下述初始配置:将交换机 S1 端口 3、交换机 S2 端口 1 和端口 2、交换机 S3 端口 3 配置为被所有 VLAN 共享的共享端口,交换机 S1 和交换机 S3 为服务器模式,交换机 S2 为客户模式。在交换机 S1 或交换机 S3 上配置 VTP 域名,如 abc。图 2.29 是完成初始配置后的交换机之间 VTP 消息交换过程。假定初始状态下,交换机只包含默认 VLAN——VLAN 1。

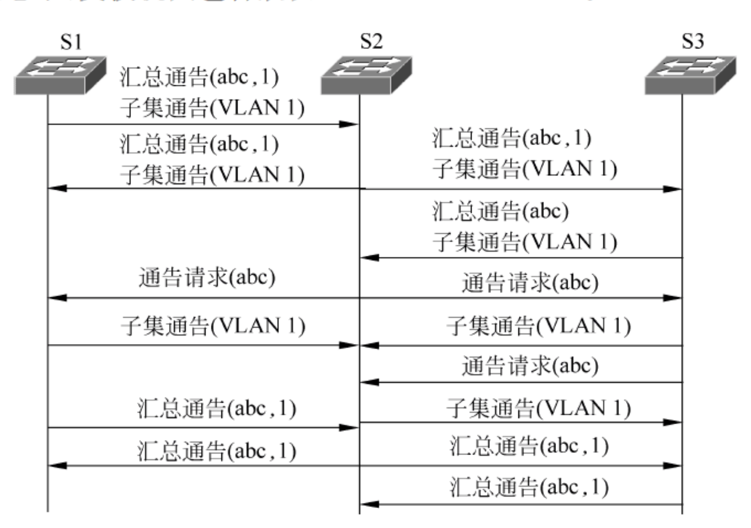


图 2.29 配置 VTP 域名过程

一旦将交换机 S1 的模式和 VTP 域名配置为服务器模式和 abc,交换机 S1 通过共享端口发送汇总通告和子集通告,子集通告中给出交换机 S1 现有的 VLAN 配置情况。交换机 S2 接收到汇总通告和子集通告后,将其 VTP 域名配置为 abc,将消息中给出的配置修订号 1 作为交换机 S1 的最新配置修订号,并根据子集通告配置 VLAN,由于子集通告中的 VLAN 情况与交换机 S2 现有的 VLAN 配置情况相同,交换机 S2 无须进行创建或删除 VLAN 操作。

交换机 S2 完成 VTP 域名配置后,立即通过共享端口发送汇总通告和子集通告,子集通告中给出交换机 S2 现有的 VLAN 配置情况。通过共享端口——端口 1 和端口 2 发送出去的汇总通告和子集通告分别到达交换机 S1 和交换机 S3。由于交换机 S1 已经配置 VTP 域名且 VTP 域名与汇总通告中的 VTP 域名相同,将消息中的配置修订号 1 作为交换机 S2 的配置修订号。由于子集通告中的 VLAN 情况与交换机 S1 现有的 VLAN 配置情况相同,交换机 S1 无须进行创建或删除 VLAN 操作。

交换机 S3 接收到汇总通告和子集通告后,将其 VTP 域名配置为 abc,将消息中给出的配置修订号 1 作为交换机 S2 的最新配置修订号,并根据子集通告配置 VLAN。由于子集

通告中的 VLAN 情况与交换机 S3 现有的 VLAN 配置情况相同,交换机 S3 无须进行创建或删除 VLAN 操作。

交换机 S3 完成 VTP 域名配置后,立即通过共享端口发送汇总通告和子集通告,子集通告中给出交换机 S3 现有的 VLAN 配置情况。通过共享端口发送出去的汇总通告和子集通告分别到达交换机 S2。由于交换机 S2 已经配置 VTP 域名且 VTP 域名与汇总通告中的 VTP 域名相同,将消息中的配置修订号 1 作为交换机 S3 的配置修订号。由于子集通告中的 VLAN 情况与交换机 S2 现有的 VLAN 配置情况相同,交换机 S2 无须进行创建或删除 VLAN 操作。

交换机 S2 和交换机 S3 在改变 VTP 域名后,通过共享端口发送通告请求,要求相邻交换机通过子集通告给出其所有 VLAN 的配置情况。由于交换机 S1、交换机 S2 和交换机 S3 的 VLAN 配置情况相同,因此,这些 VTP 消息交换过程不会改变交换机的 VLAN 配置。

在 VLAN 配置没有改变的情况下,交换机 S1、交换机 S2 和交换机 S3 周期性发送汇总通告,对于这些汇总通告,由于汇总通告中的配置修订号 1 与接收该汇总通告的交换机为该汇总通告发送者保留的配置修订号相同,接收这些汇总通告的交换机将丢弃这些汇总通告。

2）交换机 S1 创建 VLAN 2

交换机 S1 上手工创建 VLAN 2 后的 VTP 消息交换过程如图 2.30 所示。一旦在交换机 S1 上手工创建 VLAN 2,交换机 S1 的 VLAN 配置情况发生改变,交换机 S1 递增配置修订号(配置修订号变为 2),通过共享端口发送汇总通告和子集通告,子集通告中给出交换机 S1 配置的所有 VLAN 的情况。交换机 S2 接收到汇总通告和子集通告后,首先判别 VTP 域名,在确定汇总通告中给出的 VTP 域名 abc 与自己的 VTP 域名相同的情况下,继续判别配置修订号,由于汇总通告中给出的配置修订号 2 大于交换机 S2 为交换机 S1 保留的配置修订号 1,于是,将 2 作为交换机 S1 的最新配置修订号,开始处理子集通告。由于交换机 S2 没有 VLAN 2,创建 VLAN 2。由于创建 VLAN 2 的操作改变了交换机 S2 的 VLAN 配置情况,交换机 S2 递增配置修订号,通过共享端口发送汇总通告和子集通告。交换机 S1 接收到交换机 S2 发送的汇总通告和子集通告后,只是将汇总通告中的配置修订号 2 作为交换机 S1 的最新配置修订号。交换机 S3 将汇总通告中的配置修订号 2 作为交换机 S2 的最新配置修订号,创建 VLAN 2。同样,由于创建 VLAN 2 的操作改变了交换机 S3 的 VLAN 配置情况,交换机 S3 递增配置修订号,通过共享端口发送汇总通告和子集通告。交换机 S2 接收到交换机 S3 发送的汇总通告和子集通告后,只是将汇总通告中的配置修订号 2 作为交换机 S3 的最新配置修订号。

同样,在 VLAN 配置没有改变的情况下,交换机 S1、交换机 S2 和交换机 S3 周期性发送汇总通告,对于这些汇总通告,由于汇总通告中的配置修订号 2 与接收该汇总通告的交换机为该汇总通告发送者保留的配置修订号相同,接收这些汇总通告的交换机将丢弃这些汇总通告。

3）交换机 S3 创建 VLAN 3

交换机 S3 上手工创建 VLAN 3 后的 VTP 消息交换过程如图 2.31 所示。一旦在交换机 S3 上手工创建 VLAN 3,交换机 S3 的 VLAN 配置情况发生改变,交换机 S3 递增配置修订号(配置修订号变为 3),通过共享端口发送汇总通告和子集通告,子集通告中给出交换机 S3 配置的所有 VLAN 的情况。交换机 S2 接收到汇总通告和子集通告后,首先判别 VTP 域名,在汇总通告中给出的 VTP 域名 abc 与自己的 VTP 域名相同的情况下,继续判别配置修订号,由于汇总通告中给出的配置修订号 3 大于交换机 S2 为交换机 S3 保留的配置修订

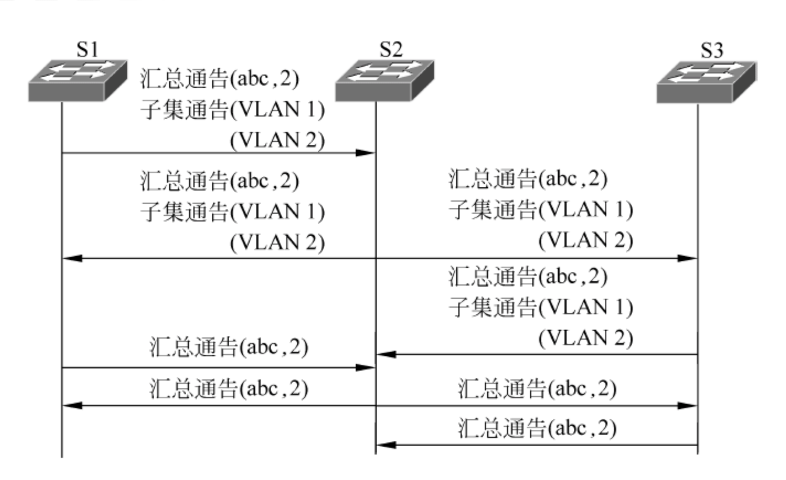


图 2.30　创建 VLAN 2 过程

号 2,将 3 作为交换机 S3 的最新配置修订号,开始处理子集通告。由于交换机 S2 没有 VLAN 3,创建 VLAN 3。由于创建 VLAN 3 的操作改变了交换机 S2 的 VLAN 配置情况,交换机 S2 递增配置修订号,通过共享端口发送汇总通告和子集通告。交换机 S1 在接收到交换机 S2 发送的汇总通告和子集通告后,创建 VLAN 3,使得交换机 S1、交换机 S2 和交换机 S3 的 VLAN 配置情况相同(每一个交换机均包含 VLAN 1、VLAN 2 和 VLAN 3)。

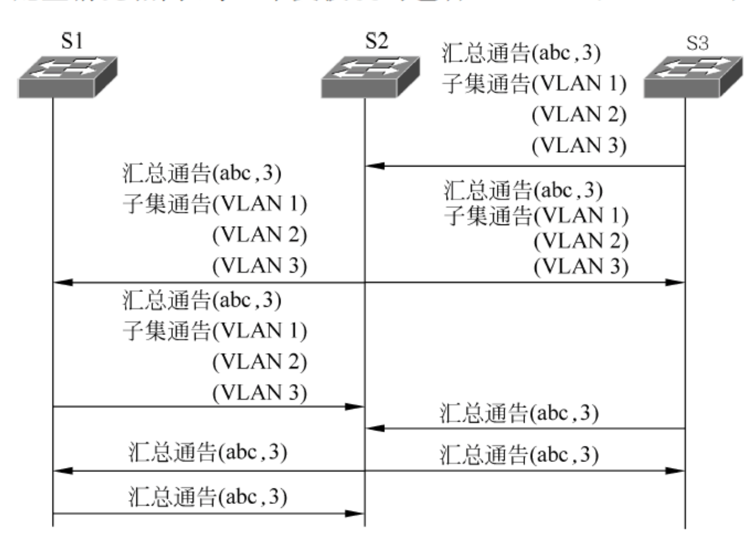


图 2.31　创建 VLAN 3 过程

4）交换机 S1 删除 VLAN 3

由于交换机 S1 和交换机 S3 的模式是服务器模式,可以在交换机 S1 和交换机 S3 上手工创建和删除 VLAN,而且任何一个服务器模式的交换机可以删除在另一个服务器模式的交换机上创建的 VLAN。因此,交换机 S1 可以删除交换机 S3 上手工创建的 VLAN 3。

在完成图 2.31 所示的 VTP 消息交换过程后,交换机 S1、交换机 S2 和交换机 S3 均包含 VLAN 1、VLAN 2 和 VLAN 3。一旦在交换机 S1 上手工删除 VLAN 3,交换机之间开始图 2.32 所示的 VTP 消息交换过程。由于交换机 S1 手工删除 VLAN 3 的操作使交换机 S1 的 VLAN 配置情况发生改变,交换机 S1 递增配置修订号(配置修订号变为 4),通过共享

端口发送汇总通告和子集通告，子集通告中给出交换机 S1 当前的 VLAN 配置情况（只存在 VLAN 1 和 VLAN 2）。交换机 S2 接收到汇总通告和子集通告后，开始根据子集通告给出的 VLAN 配置情况调整自己的 VLAN 配置，由于子集通告中没有包含 VLAN 3，交换机 S2 删除 VLAN 3。由于删除 VLAN 3 的操作改变了交换机 S2 的 VLAN 配置情况，交换机 S2 递增配置修订号，通过共享端口发送汇总通告和子集通告。交换机 S3 在接收到交换机 S2 发送的汇总通告和子集通告后，删除 VLAN 3，使得交换机 S1、交换机 S2 和交换机 S3 的 VLAN 配置情况相同（每一个交换机只包含 VLAN 1 和 VLAN 2）。

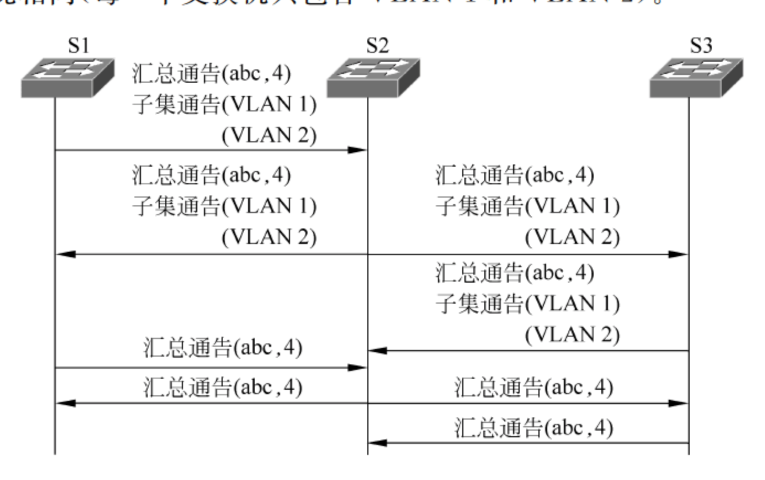


图 2.32 删除 VLAN 3 过程

需要强调的是，VTP 只是使属于相同 VTP 域的所有交换机的 VLAN 配置同步，但并不能确定交换机端口与 VLAN 之间的绑定，每一个交换机必须通过手工配置方式完成将特定交换机端口分配给特定 VLAN 的操作。

5. VTP 剪枝过程

划分 VLAN 的目的是限制广播域，使得广播帧只能在特定 VLAN 内广播，但由于 VTP 域中所有互连交换机的端口都被配置成被所有 VLAN 共享的共享端口，因此，属于任何 VLAN 的终端发送的广播帧（或以广播方式传输的单播帧）都被广播到 VTP 域内的所有交换机。如图 2.33 所示，虽然交换机 S3 及其所连接的分枝并没有连接属于 VLAN 2 的终端的接入端口，但终端 A 发送的广播帧被广播到 VTP 域内的所有交换机。

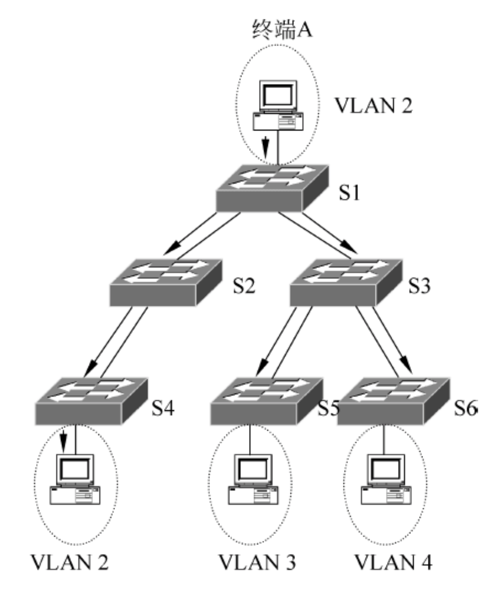


图 2.33 VLAN 2 内广播过程

为了避免图 2.33 所示的情况发生，VTP 引入了剪枝功能，如果某个服务器模式的交换机启动剪枝功能，剪枝功能将传播到 VTP 域内的所有交换机。一旦启动剪枝功能，交换机初始时除了属于默认 VLAN 的广播帧（或以广播方式传输的单播帧），共享端口并不发送属于其他 VLAN 的广播

帧。当某个交换机端口被分配给某个 VLAN,该交换机生成一个 VTP Join 消息,Join 消息中给出该 VLAN 的编号(VLAN ID)和名字,并将该 Join 消息通过所有共享端口发送出去。如果某个交换机通过某个共享端口接收到该 Join 消息,该共享端口将记录 Join 消息中给出的 VLAN 编号和名字,以后该交换机将从该共享端口发送属于该 VLAN 的广播帧。完成记录后,该交换机将从所有其他共享端口转发该 Join 消息,使该 Join 消息到达 VTP 域中的所有交换机。图 2.34 给出将交换机 S1 和交换机 S4 某个端口分配给 VLAN 2 后引发的 Join 消息传播过程,及记录 VLAN 2 的共享端口。

启动剪枝功能后,如果某个交换机接收到属于某个 VLAN 的广播帧,该交换机将从除接收该广播帧的端口以外的且符合下列条件之一的所有其他端口转发该广播帧。

- 端口为属于该 VLAN 的接入端口。
- 端口为记录了该 VLAN 编号的共享端口。

终端 A 发送的广播帧不再广播到 VTP 域内的所有交换机,而只是沿着连接属于 VLAN 2 的终端的分枝广播,如图 2.35 所示。

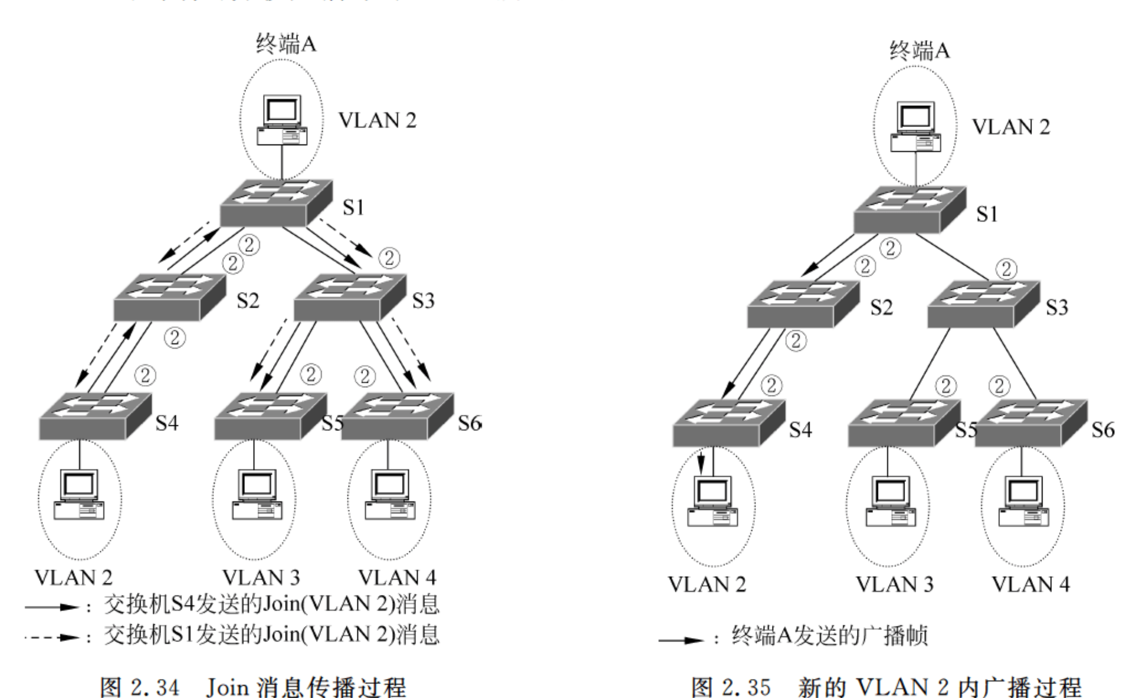


图 2.34 Join 消息传播过程

图 2.35 新的 VLAN 2 内广播过程

2.6.5 GVRP 例题解析

【例 2.4】 假定图 2.36 所示网络结构的初始配置如下:

- 交换机 S1、交换机 S2 和交换机 S3 只存在静态 VLAN——VLAN 1。
- 在交换机 S1、交换机 S2 和交换机 S3 上使能 GVRP。
- 将交换机 S1 端口 3、交换机 S2 端口 1 和端口 2、交换机 S3 端口 3 配置为被所有 VLAN 共享的共享端口。
- 在这些共享端口上使能 GVRP,并将这些共享端口的注册模式配置为 Normal 模式。

如果在交换机 S1 上手工创建静态 VLAN——VLAN 2,在交换机 S3 上手工创建静态

VLAN——VLAN 3。给出 GVRP 最终在交换机 S1、交换机 S2 和交换机 S3 上创建的 VLAN 及类型。

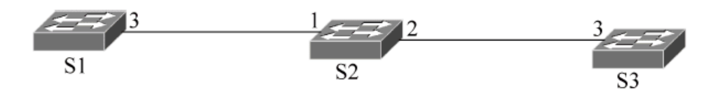


图 2.36　网络结构

【解析】 交换机 S1、交换机 S2 和交换机 S3 初始存在静态 VLAN——VLAN 1。因为交换机 S1 手工创建静态 VLAN——VLAN 2 导致交换机 S2 和交换机 S3 创建动态 VLAN——VLAN 2。因为交换机 S3 手工创建静态 VLAN——VLAN 3 导致交换机 S1 和交换机 S2 创建动态 VLAN——VLAN 3。最终得出表 2.10 所示的三个交换机的 VLAN 配置情况。

表 2.10　三个交换机的 VLAN 配置情况

交换机	VLAN	类型
S1	VLAN 1	静态
	VLAN 2	静态
	VLAN 3	动态
S2	VLAN 1	静态
	VLAN 2	动态
	VLAN 3	动态
S3	VLAN 1	静态
	VLAN 2	动态
	VLAN 3	静态

【例 2.5】 其余初始配置和例 2.4 相同，但将交换机 S1 端口 3 的注册模式配置为 Fixed 模式。如果在交换机 S1 上手工创建静态 VLAN——VLAN 2，在交换机 S3 上手工创建静态 VLAN——VLAN 3。给出 GVRP 最终在交换机 S1、交换机 S2 和交换机 S3 上创建的 VLAN 及类型。

【解析】 由于 Fixed 模式端口能够声明静态 VLAN，因此，交换机 S1 手工创建静态 VLAN——VLAN 2 依然导致交换机 S2 和交换机 S3 创建动态 VLAN——VLAN 2。由于 Fixed 模式端口不能注册动态 VLAN，因此，交换机 S3 上手工创建静态 VLAN——VLAN 3 只能导致在交换机 S2 上创建动态 VLAN——VLAN 3。最终得出表 2.11 所示的三个交换机的 VLAN 配置情况。

表 2.11　三个交换机的 VLAN 配置情况

交换机	VLAN	类型
S1	VLAN 1	静态
	VLAN 2	静态
S2	VLAN 1	静态
	VLAN 2	动态
	VLAN 3	动态

续表

交换机	VLAN	类型
S3	VLAN 1	静态
	VLAN 2	动态
	VLAN 3	静态

【例 2.6】 其余初始配置和例 2.4 相同，但将交换机 S1 端口 3 的注册模式配置为 Forbidden 模式。如果在交换机 S1 上手工创建静态 VLAN——VLAN 2，在交换机 S3 上手工创建静态 VLAN——VLAN 3。给出 GVRP 最终在交换机 S1、交换机 S2 和交换机 S3 上创建的 VLAN 及类型。

【解析】 由于 Forbidden 模式端口无法声明除 VLAN 1 以外的静态 VLAN，因此，在交换机 S1 上手工创建静态 VLAN——VLAN 2 无法导致在交换机 S2 和交换机 S3 上创建动态 VLAN——VLAN 2。由于 Forbidden 模式端口无法注册动态 VLAN，交换机 S3 上手工创建静态 VLAN——VLAN 3 只能导致在交换机 S2 上创建动态 VLAN——VLAN 3。最终得出表 2.12 所示的三个交换机的 VLAN 配置情况。

表 2.12 三个交换机的 VLAN 配置情况

交换机	VLAN	类型
S1	VLAN 1	静态
	VLAN 2	静态
S2	VLAN 1	静态
	VLAN 3	动态
S3	VLAN 1	静态
	VLAN 3	静态



习题

2.1 简述引出 VLAN 的原因，实现 VLAN 的技术基础。

2.2 802.1Q 有什么作用？连接终端的以太网交换机端口是否只能是非标记端口？为什么？

2.3 要求将图 2.37 所示网络中的终端 A、终端 B 和终端 F 划分为 VLAN 2，终端 C、终端 E 划分为 VLAN 3，终端 D、终端 G 和终端 H 划分为 VLAN 4，请给出三个以太网交换机的端口配置。给出在所有终端都广播了以自身 MAC 地址为源 MAC 地址，全 1 广播地址为目的 MAC 地址的广播帧后，3 个 VLAN 相关联的转发表内容。根据转发表内容讲述终端 A→终端 F 的传输过程，说明终端 A→终端 D 不可达的原因。

2.4 交换式以太网结构如图 3.38 所示，要求终端 A、终端 B 和终端 G 属于一个 VLAN，终端 E、终端 F 和终端 H 属于一个 VLAN，终端 C 和终端 D 属于一个 VLAN，给出交换机配置，并说明理由。

2.5 GVRP 中每当交换机增加 VLAN 属性时，该交换机发送两次 Join 消息（或是发

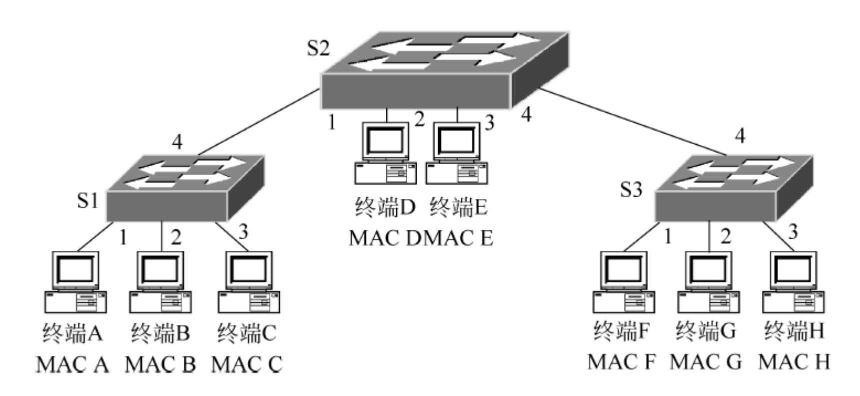


图 2.37 题 2.3 图

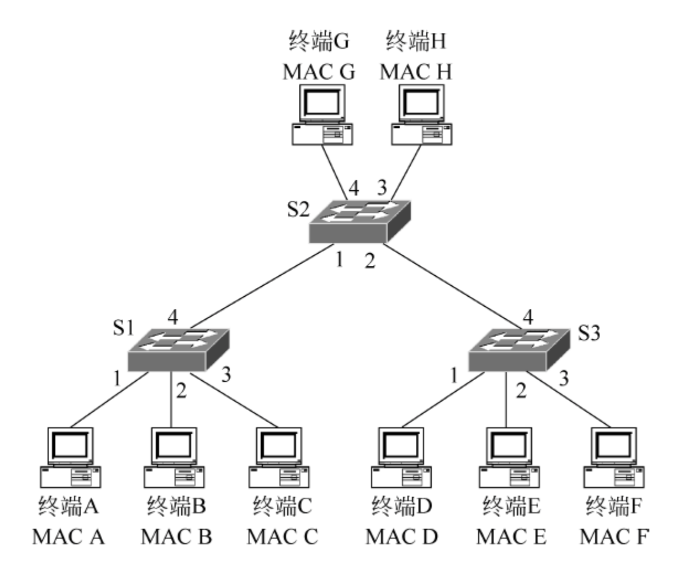


图 2.38 题 2.4 图

送一个 Join 消息后,接收一个 JoinIn 消息),是否可能发生因为两个 Join 消息丢失,导致其他交换机永远无法完成相应 VLAN 属性注册的问题?

2.6 GVRP 中每当交换机删除属性时,该交换机发送一次 Leave 消息,是否可能发生因为 Leave 消息丢失,导致其他交换机永远无法完成相应 VLAN 属性注销的问题?

2.7 假定图 2.39 所示的网络结构的初始配置如下:

- 交换机 S1、交换机 S2 和交换机 S3 只存在静态 VLAN——VLAN 1。
- 在交换机 S1、交换机 S2 和交换机 S3 上使能 GVRP。
- 将交换机 S1 端口 3、交换机 S2 端口 1 和端口 2、交换机 S3 端口 3 配置为被所有 VLAN 共享的共享端口。
- 交换机 S1 端口 3 的注册模式为 Fixed 模式,交换机 S2 端口 1 的注册模式为 Forbidden 模式,其他共享端口的注册模式配置为 Normal 模式。

如果在交换机 S1 上手工创建静态 VLAN——VLAN 2,在交换机 S3 上手工创建静态 VLAN——VLAN 3。给出 GVRP 最终在交换机 S1、交换机 S2 和交换机 S3 上创建的 VLAN 及类型。

图 2.39 题 2.7 图

2.8 简述交换机发送通告请求的条件，针对每一种条件给出实例。

2.9 交换式以太网结构如图 2.40 所示，不同填充图案的端口属于不同的 VLAN，所有端口为非标记端口(Access 端口)，给出所有连接在不同交换机上且能够实现通信的终端对，并简述原因。

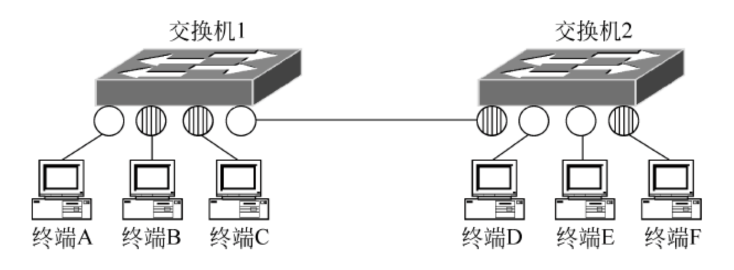


图 2.40 题 2.9 图

2.10 交换式以太网结构如图 2.41 所示，不同填充图案的端口属于不同的 VLAN，交换机 1 端口 4 和交换机 2 端口 1 为被所有 VLAN 共享的共享端口，且为 802.1Q 标记端口，其他端口为非标记端口(Access 端口)，给出所有连接在不同交换机上且能够实现通信的终端对，并简述结果与题 2.9 不同的原因。

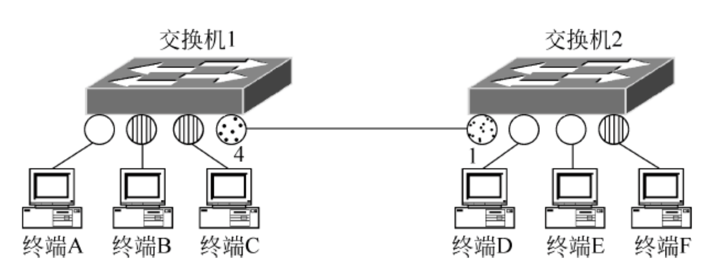


图 2.41 题 2.10 图

2.11 交换机连接终端和集线器方式及端口分配给各个 VLAN 的情况如图 2.42 所示，假定终端后面的字符表示终端的 MAC 地址，初始转发表为空表，回答以下问题。

① 终端 A 发送的目的 MAC 地址为 B 的 MAC 帧到达哪些终端？

② 终端 B 发送的目的 MAC 地址为 A 的 MAC 帧到达哪些终端？

③ 终端 E 发送的目的 MAC 地址为 B 的 MAC 帧到达哪些终端？

④ 终端 B 发送的目的 MAC 地址为 E 的 MAC 帧到达哪些终端？

⑤ 终端 B 发送的目的 MAC 地址为广播地址的 MAC 帧到达哪些终端？

⑥ 终端 F 发送的目的 MAC 地址为 E 的 MAC 帧到达哪些终端？

2.12 引入专用 VLAN 的原因是什么？简述 Cisco 专用 VLAN 技术的实现要点。

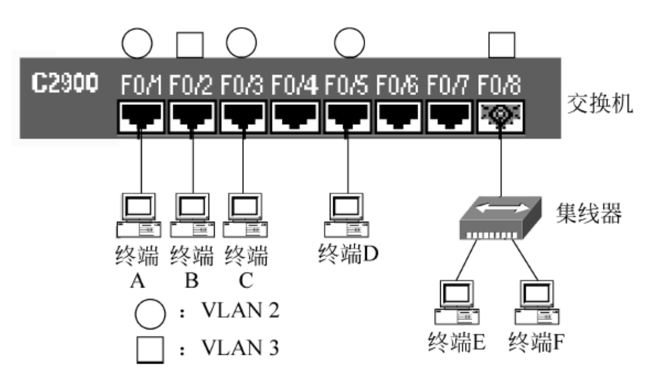

C2900
F0/1 F0/2 F0/3 F0/4 F0/5 F0/6 F0/7 F0/8
交换机
集线器
终端A
终端B
终端C
终端D
终端E
终端F
：VLAN 2
：VLAN 3

图 2.42 题 2.11 图

第3章 生成树协议

交换机工作原理要求交换机之间不允许存在环路，但树型结构交换式以太网的可靠性存在问题，一旦网络中某段链路或是某个交换机发生故障，会导致一部分终端无法和网络中的其他终端通信。生成树协议允许设计一个存在冗余链路的网络，但在网络运行时，通过阻塞某些端口使整个网络没有环路。当某条链路或是某个交换机发生故障时，通过重新开通原来阻塞的一些端口，使网络终端之间依然保持连通性，而又没有形成环路，这样，既提高了网络的可靠性，又消除了环路带来的问题。

3.1 生成树协议的作用

3.1.1 环路引发广播风暴

网桥在没有完全建立转发表之前，以广播方式转发 MAC 帧的机制对网桥之间的连接方式带来很大限制，图 3.1 是两个网桥之间存在环路的连接方式，这种连接方式会对 MAC 帧传输带来一些问题。

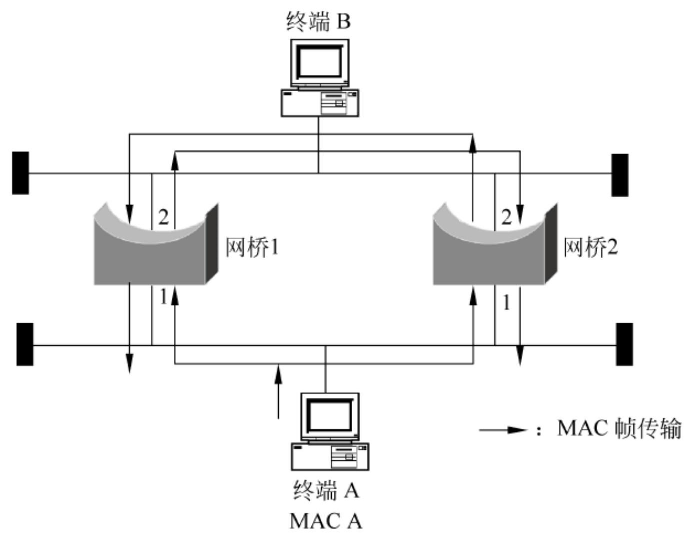


图 3.1　网桥之间存在环路的连接方式

假定网桥 1 和网桥 2 中的转发表还没有学习到终端 B 的 MAC 地址，当终端 A 向终端 B 发送 MAC 帧时，网桥 1 的端口 1 和网桥 2 的端口 1 均收到该 MAC 帧，由于网桥 1 和网

桥 2 的转发表中均没有和终端 B 的 MAC 地址匹配的转发项，网桥 1 和网桥 2 又都从各自的端口 2 将该 MAC 帧发送出去（广播方式）。同样，网桥 1 端口 2 通过争用共享媒体发送的 MAC 帧被网桥 2 的端口 2 收到，而网桥 2 端口 2 通过争用共享媒体发送的 MAC 帧又被网桥 1 的端口 2 收到。此时，虽然终端 B 已经重复两次收到该 MAC 帧，但网桥 1 和网桥 2 仍然又通过端口 1 将该 MAC 帧发送出去（广播方式）。使得该 MAC 帧在由网桥 1、网桥 2 构成的环路内不停地兜圈子，白白浪费了网络带宽。导致该问题发生的罪魁祸首就是网桥之间存在的环路，环路引发广播风暴。

环路除了引发广播风暴外，还会造成转发表的错误，当图 3.1 中网桥 1 和网桥 2 通过端口 1 接收到终端 A 发送的 MAC 帧时，网桥 1 和网桥 2 在转发表中建立将终端 A 的 MAC 地址（MAC A）与端口 1 绑定在一起的转发项。但当网桥 1 和网桥 2 再次通过端口 2 接收到该 MAC 帧时，将转发项中与 MAC A 绑定的端口由端口 1 改变为端口 2，如果此时终端 B 向终端 A 发送 MAC 帧，网桥 1 和网桥 2 将因为该 MAC 帧的输入端口与转发表指定的输出端口相同而丢弃该 MAC 帧。当然，随着终端 A 发送的 MAC 帧不断地在由网桥 1、网桥 2 构成的环路内兜圈子，网桥 1 和网桥 2 转发表中与 MAC A 匹配的转发项的转发端口也在不断地发生变化。

3.1.2 树型网络的弱可靠性

为了消除因为环路引发的广播风暴，用网桥互连而成的网络中，任何两个终端之间只允许存在一条传输路径。在设计网络时做到这一点并不难，可以设计一个树型结构的网络，终端为树的叶结点，从树根到任何叶结点之间不容许有任何环路存在（只允许有一条传输路径），这样的树型结构网络如图 3.2 所示。但这种网络结构的可靠性不高，任何一段链路发生故障，就有可能使一部分终端无法和网络中的其他终端通信。图中网桥 2 连接网桥 3 的链路一旦发生故障，网桥 3 连接的终端将无法和网桥 1 连接的终端通信。因此，树型结构网络的可靠性不高。

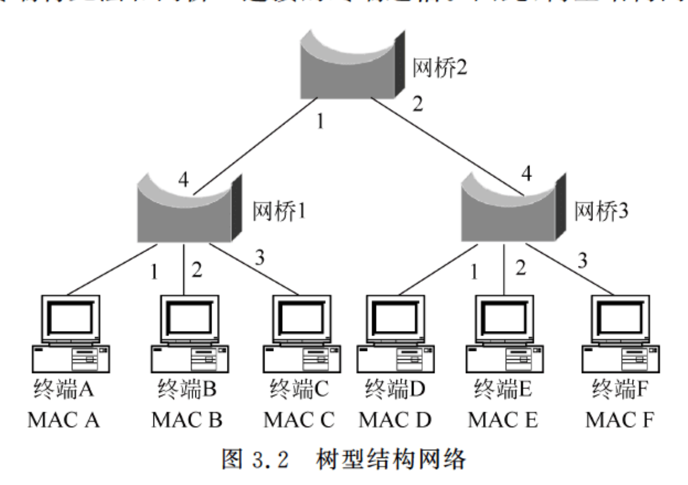


图 3.2　树型结构网络

3.1.3 生成树协议的由来和发展

是否能够设计这样一种网络，它存在冗余链路，但在网络运行时，通过阻塞某些端口使整个网络没有环路，当某条链路因为故障无法通信时，通过重新开通原来阻塞的一些端口，

使网络终端之间依然保持连通性，而又没有形成环路，这样，既提高了网络的可靠性，又消除了环路带来的问题。生成树协议(Spanning Tree Protocol，STP)就是这样一种机制，图 3.3 就是描述生成树协议作用过程的示意图。

原始网络结构如图 3.3(a)所示，网桥之间存在环路，以此提高网络的可靠性。STP 阻塞形成环路的端口后，网络结构变成图 3.3(b)所示的以根网桥为树根的树型结构。但一旦网桥之间链路发生故障，如图 3.3(c)所示的网桥 4 和网桥 5、网桥 5 和网桥 7 之间链路发生故障，STP 通过重新开通原来阻塞的一些端口，使网桥之间依然保持连通性，如图 3.3(d)所示。

STP 经过不断发展，衍生出快速生成树协议(Rapid Spanning Tree Protocol，RSTP)和多生成树协议(Multiple Spanning Tree Protocol，MSTP)。

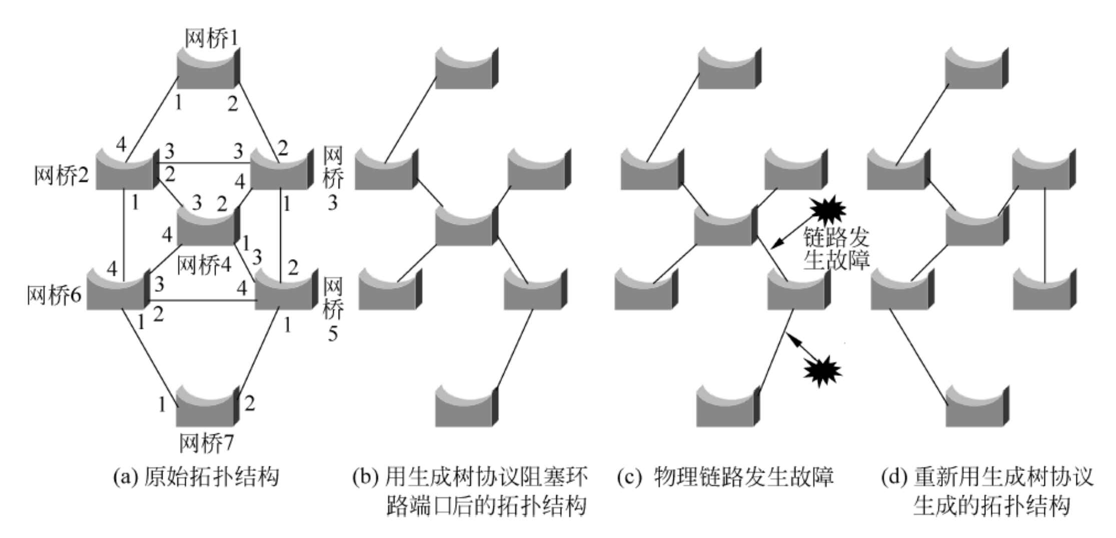


图 3.3 生成树协议作用过程示意图

3.2 生成树协议工作过程

3.2.1 生成树协议操作步骤

为了将图 3.3(a)所示的网状拓扑结构变成图 3.3(b)所示的以网桥 4 为根的树型拓扑结构，生成树协议必须完成下述步骤：

(1) 产生根网桥；

(2) 找出其他网桥与根网桥之间路径最短的根端口；

(3) 对任何和两个或以上网桥端口相连的链路，找出指定网桥和指定端口。

根网桥是形成图 3.3(b)所示的生成树后，作为树根的网桥。而某个网桥的根端口是指这样一个端口：网桥通过该端口到达根网桥的路径最短。某个网桥的根路径距离就是通过该网桥的根端口到达根网桥的距离，它是该网桥到达根网桥的最短路径距离。如果某条链路不是根路径所经过的链路，且这条链路连接两个或两个以上网桥端口，该链路就会形成环路，这样的链路只能由一个端口进行正常的输入/输出操作，其余端口都将被阻塞，进行正常的输入/输出操作的端口就是该链路的指定端口，端口所在的网桥就是指定网桥。为了产生

根网桥，必须对所有网桥分配一个标识符，标识符格式如图 3.4 所示，前 2 个字节的网桥优先级可以手工配置，后 6 个字节的网桥 MAC 地址是厂家在生产网桥时设定的，不能修改。所有网桥中网桥标识符值最小的网桥为根网桥。因此，如果希望某个网桥成为根网桥，可将该网桥的网桥优先级字段配置成较小的值。

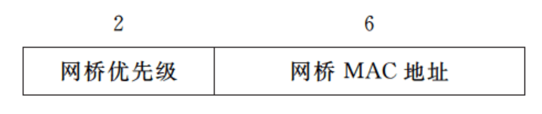


图 3.4 网桥标识符

为了计算根路径距离，必须为网桥的每一端口设置路径距离，端口的路径距离称为端口路径距离。早期生成树协议规定端口路径距离＝1000Mb/s/端口速率，在端口速率为 10Mb/s、100Mb/s 时，用这个公式计算出的端口路径距离不仅是整数，而且能够反映出端口的速率差别，即端口速率越高，路径距离越小，端口速率差 10 倍，则路径距离也差 10 倍。但当端口速率为 10Gb/s 时，求得的端口路径距离为小数，这和生成树协议对端口路径距离必须为整数的要求不符。如果将所有小于 1 的端口路径距离置为 1，又混淆了端口速率差别，因此，IEEE 重新对端口速率和端口路径距离之间的对应关系作了定义，如表 3.1 所示。

表 3.1 端口速率和端口路径距离之间的对应关系

端口速率	端口路径距离
10Mb/s	100
100Mb/s	19
1Gb/s	4
10Gb/s	2

网桥 A 至网桥 B 的路径距离就是网桥 A 至网桥 B 传输路径所经过的所有输入端口的端口路径距离之和。如图 3.3(a)所示的网桥 4 至网桥 7 的其中一条传输路径(网桥 4→网桥 5→网桥 7)的路径距离等于网桥 5 端口 3 的端口路径距离＋网桥 7 端口 2 的端口路径距离。

3.2.2 生成树协议构建生成树过程

1. BPDU 格式

网桥通过相互交换网桥协议数据单元(Bridge Protocol Data Unit，BPDU)来学习网络拓扑结构，构建生成树。BPDU 只在直接连接的两个网桥之间传输，不能转发。网桥协议数据单元(BPDU)格式及主要包含的内容如图 3.5 所示。BPDU 的目的 MAC 地址固定为 01:80:C2:00:00:00，网桥将目的 MAC 地址为 01-80-C2-00-00-00 的 MAC 帧提交生成树协议进程处理。

根网桥标识符是发送该 BPDU 的网桥学习到的根网桥的网桥标识符，任何一个网桥都将通过接收到的 BPDU 学习到的网桥标识符最小的网桥作为根网桥。

根路径距离是发送该 BPDU 的网桥至根网桥的最短路径的距离。

发送网桥标识符是发送该 BPDU 的网桥的网桥标识符。

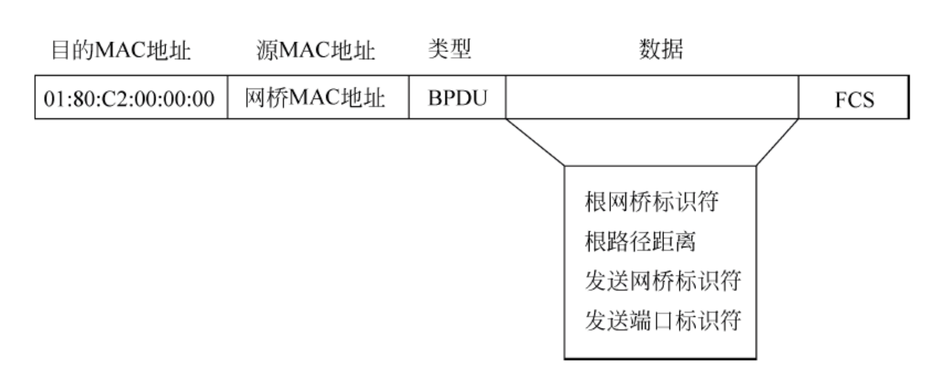


图 3.5　BPDU 格式

发送端口标识符是发送该 BPDU 的网桥输出该 BPDU 的端口的端口标识符。

2. 生成树协议工作原理

1）确定根网桥、根路径距离和根端口

每一个网桥加电或初始化后，认为自身就是根网桥，并以此为每一个端口产生端口 BPDU，所有端口周期性地发送端口 BPDU。假定某个网桥的标识符为 I(简称网桥 I)，加电或初始化后，该网桥确定根网桥标识符＝I，根路径距离＝0，并以此为所有端口产生端口 BPDU，不同端口的端口 BPDU 中根网桥标识符、根路径距离和发送网桥标识符都是相同的，不同的只是端口标识符，如端口标识符为 P_i 的端口 BPDU 为＜根网桥标识符＝I，根路径距离＝0，发送网桥标识符＝I，发送端口标识符＝P_i＞，简写为＜I，0，I，P_i＞，端口 P_i 周期性发送 BPDU＜I，0，I，P_i＞。需要强调的是，只有根网桥周期性地通过所有非阻塞端口发送端口 BPDU，某个网桥一旦确定自身不是根网桥，就不再自发地周期性地通过非阻塞端口发送端口 BPDU。如果某个网桥的根网桥和根路径距离通过某个接收到的 BPDU，或自身 BPDU 得出，该 BPDU 称为该网桥的网桥最佳 BPDU，显然，初始化后，网桥 I 的网桥最佳 BPDU 为该网桥自身 BPDU＜I，0，I＞。

初始化后，当网桥 I 通过端口标识符为 P 的端口接收到 BPDU＜X，N，Y，P_j＞（该 BPDU 由网桥 Y 通过端口 P_j 发出，而且表明网桥 Y 确定网桥 X 为根网桥，网桥 Y 到达根网桥 X 的根路径距离为 N），Z＝N＋端口 P 的端口路径距离，如果 X＜I，进行如下操作：

（1）根网桥标识符＝X，根路径距离＝Z，根端口标识符＝P，BPDU＜X，N，Y，P_j＞为网桥最佳 BPDU；

（2）重新为每一个端口生成端口 BPDU，端口标识符为 P_i 的端口 BPDU 为＜X，Z，I，P_i＞；

（3）一旦确定根网桥不是自身，只有在通过根端口接收到网桥最佳 BPDU 的情况下，才通过除根端口以外的每一个非阻塞端口发送端口 BPDU。

一般情况下，假定网桥 I 根据目前的网桥最佳 BPDU ＜X，N1，Y1，P_1＞得出根网桥标识符 X，根路径距离 Z，根端口标识符 P。当网桥从端口标识符为 P_i 的端口接收到 BPDU＜X1，N2，Y2，P_j＞，Z2＝N2＋端口 P_i 的端口路径距离，依次进行下述各项操作，在上一项条件不成立的情况下，进行下一项操作：

① IF X1＜X，X＝X1，Z＝Z2，P＝P_i；

② IF X1＝X · AND · Z2＜Z，Z＝Z2，P＝P_i；

③ IF X1＝X · AND · Z2＝Z · AND · N2＜N1，P＝P_i；

④ IF X1＝X · AND · Z2＝Z · AND · N2＝N1 · AND · Y2＜Y1，P＝P_i；

⑤ IF X1＝X · AND · Z2＝Z · AND · N2＝N1 · AND · Y2＝Y1 · AND · P_i＜P，P＝P_i。

在其中一项条件成立的情况下，将 BPDU＜X1，N2，Y2，P_j＞作为网桥最佳 BPDU，得出新的根网桥标识符 X＝X1，根路径距离 Z＝Z2，根端口标识符 P＝P_i。

一旦根网桥或根端口发生改变，网桥 I 重新为每一个端口生成端口 BPDU，如端口标识符为 P_k 的端口 BPDU 为＜X，Z，I，P_k＞，除根端口以外，每一个非阻塞端口发送端口 BPDU。以后，网桥只有在通过根端口接收到网桥最佳 BPDU 的情况下，才通过除根端口以外的每一个非阻塞端口发送端口 BPDU。

确定网桥最佳 BPDU 的过程就是找到一条通往标识符最小的网桥的最短路径的过程，如果 BPDU＜X1，N2，Y2，P_j＞为网桥最佳 BPDU，表明网桥 Y2 是网桥 I 通往网桥 X1 的最短路径的上游网桥，网桥 I 通往网桥 X1 的最短路径距离＝网桥 Y2 通往网桥 X1 的最短路径距离 N2＋连接上游网桥的链路的路径距离。比较两个 BPDU 哪一个更优，就是比较：①以发送 BPDU 的网桥为上游网桥的两条路径中哪一条是通往标识符最小的网桥的路径；②如果两条路径通往同一个网桥，哪一条路径的距离更短；③如果两条路径距离相等，网桥 I 连接哪一条路径的端口的传输速率较快；④如果连接两条路径的端口的速率相同，哪一条路径的上游网桥的网桥标识符较小；⑤如果两条路径有着同一个上游网桥，哪一条路径连接的端口的端口标识符较小。

2）确定阻塞端口

假定网桥 I 得出的根网桥标识符为 X，根路径距离为 Z，根端口标识符为 P。当网桥 I 从端口标识符为 P_i 的端口接收到 BPDU＜X1，N1，Y1，P_j＞，在确定 BPDU＜X1，N1，Y1，P_j＞不是网桥最佳 BPDU 且 X＝X1 的前提下，依次进行下述各项操作，在上一项条件不成立的情况下，进行下一项操作：

① IF N1＜Z，阻塞端口 P_i；

② IF N1＝Z · AND · Y1＜I，阻塞端口 P_i；

③ IF N1＝Z · AND · Y1＝I · AND · P_j＜P_i，阻塞端口 P_i。

某个端口为阻塞端口，表明通过该端口接收到的 BPDU 优于该端口的端口 BPDU，如上例中，端口 P_i 的端口 BPDU 为＜X，Z，I，P_i＞，如果通过端口 P_i 接收到 BPDU＜X1，N1，Y1，P_j＞，依次比较：Z 和 N1；I 和 Y1；P_i 和 P_j。在前一项相等的情况下，比较下一项，只要其中一项比较结果是接收到的 BPDU 中的值小于端口 BPDU 的值，表明接收到的 BPDU 优于端口 BPDU，阻塞该端口，并将接收到的 BPDU 作为该端口的端口最佳 BPDU。一旦某个端口处于阻塞状态，只允许接收 BPDU。需要指出的是：确定阻塞端口的前提是接收到的 BPDU 不是网桥最佳 BPDU，因此，该 BPDU 不会改变网桥中已经确定的根网桥标识符和根路径距离。当一条链路连接两个或两个以上端口，且这些端口位于不同网桥时，如果这些网桥的根路径距离不同，确定根路径距离最小的网桥为指定网桥，位于指定网桥的端口为指定端口。如果多个端口所在网桥的根路径距离相同，选择标识符较小的网桥为指定网桥，位于指定网桥的端口为指定端口。当多个端口位于同一网桥时，选择标识符较小的端口为指定端口。

如果网桥新接收到的 BPDU 改变了网桥的根网桥标识符或根路径距离，即网桥最佳

BPDU发生改变，则需要重新更新每一个端口的端口BPDU，如果更新后的端口BPDU优于导致该端口阻塞的端口最佳BPDU，则重新将端口BPDU作为该端口的端口最佳BPDU，将该端口从阻塞状态转变为非阻塞状态。

在生成树协议收敛之前，端口状态是不断变化的，因此，只能将持续一段时间处于非阻塞状态的端口转换成转发端口，要求的持续时间是生成树协议的收敛时间。只允许转发端口正常输入/输出MAC帧。每一个网桥转换成的转发端口中只有一个是根端口，其余为指定端口。

3）BPDU最大传输时延

在STP收敛过程中可能存在BPDU传输环路，有可能造成BPDU在网络中无休止地转发。为了防止这种情况发生，BPDU中增加了BPDU传输时延（Message Age）字段，BPDU在该字段中累计经过链路传输的时延和触发非根网桥发送端口BPDU所需要的时延，一旦累计时延超过BPDU最大存活时间（Max Age），网桥就丢弃该BPDU，不再由该BPDU触发端口BPDU的传输过程。

3. 生成树协议操作实例

假定图3.3中网桥i的MAC地址为ii:ii:ii:ii:ii:ii，所有网桥端口速率均为100Mb/s，根据表3.1求得每一个端口的端口路径距离为19。网桥4的优先级为100，其余网桥的优先级为200，网桥4的网桥标识符简写为104，其余网桥的网桥标识符简写为20i，i是图3.3中网桥的编号。网桥5初始化后确定自身为根网桥，网桥最佳BPDU为＜205,0,205＞，各个端口的端口BPDU分别为＜205,0,205,i＞，i为端口号，周期性地通过各个端口发送端口BPDU。

一旦网桥5通过端口2接收到网桥3发送的BPDU＜203,0,203,1＞，将端口2的端口路径距离累加到BPDU中的根路径距离，得出累加后的值为19，由于根网桥标识符一项的比较结果是接收到的BPDU优于网桥最佳BPDU(203＜205)，网桥5确定网桥3为根网桥，端口2为根端口，求出根路径距离为19，确定BPDU＜203,0,203,1＞为网桥最佳BPDU，重新得出各个端口的端口BPDU分别为＜203,19,205,i＞，i为端口号。网桥5只有在通过根端口接收到网桥最佳BPDU的情况下，才通过除根端口以外的所有其他端口发送端口BPDU。

一旦网桥5通过端口3接收到网桥4发送的BPDU＜104,0,104,1＞，将端口3的端口路径距离累加到BPDU中的根路径距离，得出累加后的值为19，同样由于根网桥标识符一项的比较结果是接收到的BPDU优于网桥最佳BPDU(104＜203)，网桥5确定网桥4为根网桥，端口3为根端口，求出根路径距离为19，确定BPDU＜104,0,104,1＞为网桥最佳BPDU，再次得出各个端口的端口BPDU分别为＜104,19,205,i＞，i为端口号。网桥5只有在通过根端口接收到网桥最佳BPDU的情况下，才通过除根端口以外的所有其他端口发送端口BPDU。

假定网桥6从网桥4接收到BPDU＜104,0,104,4＞，得出各个端口的端口BPDU分别为＜104,19,206,i＞，i为端口号。如果网桥6通过端口1和端口2发送的端口BPDU（＜104,19,206,1＞和＜104,19,206,2＞）分别被网桥7和网桥5接收，网桥7将其作为网桥最佳BPDU。网桥5从端口4接收BPDU＜104,19,206,2＞后，首先确定它不是网桥最

佳 BPDU，因为，将端口路径距离累加到该 BPDU 根路径距离后的结果是 38，大于根据网桥最佳 BPDU＜104，0，104，1＞求出的根路径距离 19。其次，用该 BPDU 和网桥 5 端口 4 的端口 BPDU 比较，由于发送网桥标识符一项的比较结果是端口 BPDU 优于网桥 6 发送的 BPDU（205＜206），网桥 5 端口 4 维持非阻塞状态不变。

网桥 5 分别从端口 1 和端口 4 发送端口 BPDU（＜104，19，205，1＞和＜104，19，205，4＞），网桥 7 从端口 1 接收 BPDU＜104，19，205，1＞，发现发送网桥标识符一项的比较结果是该 BPDU 优于网桥最佳 BPDU＜104，19，206，1＞（205＜206），网桥 7 将其作为网桥最佳 BPDU，并得出根网桥标识符＝104，根路径距离＝38，根端口＝端口 2。网桥 6 从端口 2 接收 BPDU ＜104，19，205，4＞后，首先确定它不是网桥最佳 BPDU，因为，将端口路径距离累加到该 BPDU 根路径距离后的结果是 38，大于根据最佳 BPDU＜104，0，104，4＞求出的根路径距离 19。其次，用该 BPDU 和网桥 6 端口 2 的端口 BPDU 比较，由于发送网桥标识符一项的比较结果是该 BPDU 优于网桥 6 端口 2 的端口 BPDU（205＜206），网桥 6 端口 2 成为阻塞端口，该 BPDU 成为网桥 6 端口 2 的端口最佳 BPDU。

一旦生成树协议收敛，所有网桥端口不是成为阻塞端口，就是转换成转发端口（根端口和指定端口），图 3.3（b）就是删除连接阻塞端口链路后得到的网络拓扑结构。

从图 3.3（b）中看到，网桥 7 端口 1 和网桥 6 端口 1 之间的链路被删除了，实际上网桥 6 端口 1 是转发端口，如果网桥 6 端口 1 和网桥 7 端口 1 之间链路通过集线器连接终端，如图 3.6（a）所示，则运行生成树协议后得到的树型结构如图 3.6（b）所示。

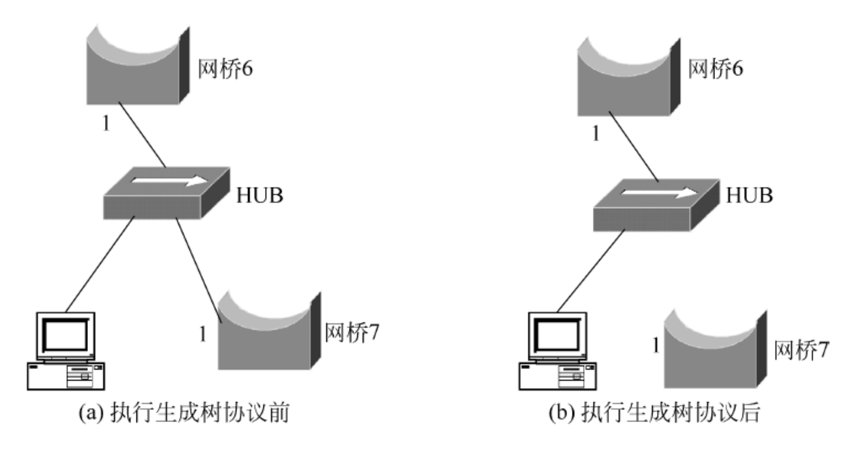


图 3.6　生成树协议作用过程示意图

3.2.3　生成树协议的容错功能

网桥中无论是网桥最佳 BPDU，还是每一个端口的端口最佳 BPDU 都设置了定时器，为了保证网桥最佳 BPDU 和每一个端口的端口最佳 BPDU 有效，必须在定时器溢出前通过根端口接收到网桥最佳 BPDU，通过对应端口接收到端口最佳 BPDU。如果网桥最佳 BPDU 对应的定时器溢出，网桥重新回到初始化状态，通过比较各个端口接收到的 BPDU，找出网桥最佳 BPDU。如果端口最佳 BPDU 对应的定时器溢出，该端口回到非阻塞状态，将端口 BPDU 作为端口最佳 BPDU。假定发生图 3.3（c）所示的链路故障，由于网桥 5 无法通过端口 3 接收到网桥最佳 BPDU＜104，0，104，1＞，导致该 BPDU 对应的定时器溢出，网

桥5设定BPDU＜205,0,205＞为网桥最佳BPDU,通过比较分别从端口2和端口4接收到的BPDU(＜104,19,203,1＞和＜104,19,206,2＞),发现发送网桥标识符一项的比较结果是网桥3发送的BPDU优于网桥6发送的BPDU(203＜206),网桥5端口2成为根端口,并以此得出端口4的端口BPDU为＜104,38,205,4＞。由于网桥6发送的BPDU＜104,19,206,2＞优于网桥5端口4的端口BPDU,端口4成为阻塞端口,BPDU＜104,19,206,2＞为网桥5端口4的端口最佳BPDU。同理,网桥6端口2成为指定端口。同样,网桥7将网桥6发送的BPDU作为网桥最佳BPDU,端口1成为根端口,网络拓扑结构转变为图3.3(d)所示。

3.2.4 端口状态和定时器

如果网络拓扑结构发生变化,有些端口可能需要从阻塞状态转变为转发状态,有些端口可能需要从转发状态转变为阻塞状态。在STP收敛过程中,如果一些需要从转发状态转变为阻塞状态的端口仍维持为转发状态,而另一些需要从阻塞状态转变为转发状态的端口已经完成状态转变,就会导致环路。反之,如果一些需要从转发状态转变为阻塞状态的端口还没有完成状态转变,而另一些需要从阻塞状态转变为转发状态的端口仍维持阻塞状态,就会导致一部分终端短暂地无法和网络中的其他终端通信。发生后一种情况的后果不是十分严重,随着STP收敛,终端之间的连通性会得到恢复。发生前一种情况的后果十分严重,广播风暴会使得网络无法实现终端之间的正常通信。

为避免在STP收敛过程中出现环路的情况,STP一是设置了3个定时器,二是使端口具有5种不同的状态,而且规定了图3.7所示的端口状态迁移过程。3个定时器分别是BPDU最大存活时间定时器、间隔时间定时器和转发时延定时器。

BPDU最大存活时间定时器(Max Age):时间初值为手工配置的网桥最佳BPDU和端口最佳BPDU的最大存活时间,每当接收到或生成网桥最佳BPDU和端口最佳BPDU时,复位该定时器。一旦定时器溢出,表示原有的网桥最佳BPDU和端口最佳BPDU无效。由此表明,在最大存活时间所规定的时间段内必须再次接收或重新生成网桥最佳BPDU和端口最佳BPDU,否则,将使网桥或端口进入初始状态。网桥初始状态将以网桥自身为根网桥,并因此开始STP操作过程。端口初始状态以根据网桥最佳BPDU推导出的端口BPDU作为端口最佳BPDU。

间隔时间定时器(Hello Time):时间初值为手工配置的根网桥从所有非阻塞端口发送BPDU的时间间隔。每当该定时器溢出,根网桥通过所有非阻塞端口发送端口BPDU。

转发时延定时器(Forward Delay):时间初值是手工配置的BPDU从根网桥传播到最外围网桥所需要的时间。一旦定时器溢出,将导致端口状态发生迁移,如图3.7所示。

三个定时器初值均有默认值,Max Age是20s,Hello Time是2s,Forward Delay是15s。

表3.2给出了端口五种状态下所具有的能力。图3.7给出了端口状态迁移过程和引发端口状态发生转变的条件。关闭状态表示端口不能工作,可以通过命令关闭或开启端口。网桥初始状态时将自己定义为根网桥,根网桥所有端口初始时处于侦听状态。其他处于关闭状态的端口,通过开启命令开启后,进入阻塞状态。无论端口处于何种状态,一旦用命令关闭该端口,该端口立即进入关闭状态。一旦STP确定某个端口属于下述三种端口类型之

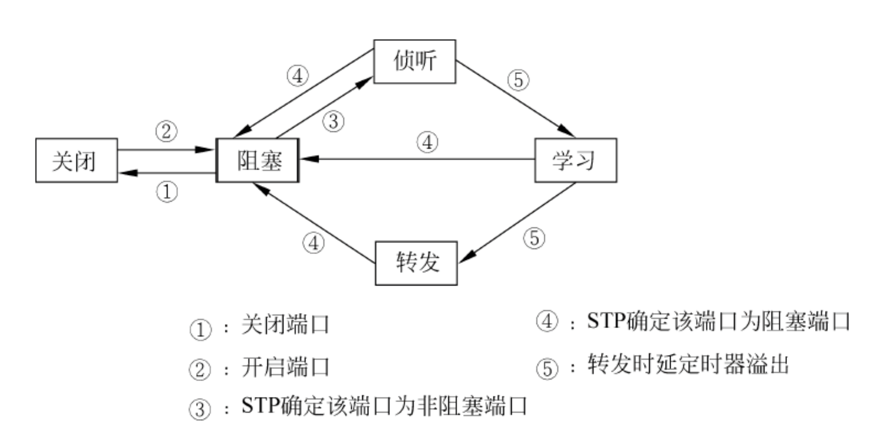


图 3.7　端口状态迁移图

一,那么该端口为非阻塞端口,端口进入侦听状态。

① 为根网桥的端口且该端口没有进入阻塞状态;

② 非根网桥的根端口;

③ 非根网桥的指定端口。

一旦某个端口进入侦听状态,则启动转发时延定时器;一旦转发时延定时器溢出,端口进入学习状态,就再次启动转发时延定时器;一旦转发时延定时器溢出,端口进入转发状态,只有处于转发状态的端口才能正常输入/输出终端之间的 MAC 帧(称为数据帧,以便区别 BPDU 这样的控制帧)。如果 STP 确定某个端口不属于上述三种端口类型,那么该端口为阻塞端口,端口立即进入阻塞状态。从中可以看出,当网络拓扑结构发生变化,某个端口从阻塞状态转变为转发状态需要经过 Max Age+2×Forward Delay 时间,这样设计的目的是保证在 STP 收敛过程中,不会发生数据帧传输路径形成环路的情况。但可能使网络的连通性短时间出现问题。

表 3.2　端口状态及能力

端口状态	端口能力
关闭(Disabled)	不能收发任何类型 MAC 帧
阻塞(blocking)	不能收发数据帧,能够接收 BPDU
侦听(listening)	不能收发数据帧,能够收发 BPDU
学习(Learning)	不能收发数据帧,能够收发 BPDU,并学习地址
转发(Forwarding)	能够正常收发 MAC 帧,并学习地址

3.2.5　网桥转发表刷新机制

1. 生成树结构与转发表内容的一致性问题

当如图 3.8(a)所示的原始网络结构通过 STP 成功构建如图 3.8(b)所示的以网桥 B3 为根网桥的生成树,且所有没有处于阻塞和关闭状态的端口均已转变为转发状态,端口开始正常输入/输出终端之间的数据帧,并因此通过地址学习建立如图 3.8 所示的转发表。一旦网络物理结构发生变化,将重新通过 STP 构建新的生成树,从网络物理结构发生变化,到通

过 STP 成功构建新的生成树所需要的时间大约为 Max Age＋2×Forward Delay，根据默认值计算所得的结果是 50s。

虽然如图 3.8(c)所示的新构建的生成树与如图 3.8(b)所示的旧的生成树相差很大，但网桥的转发表不会因此自动改变，如果网桥转发表中一部分转发项维持不变，就无法通过新构建的生成树实现终端之间的通信，如果 B5 转发表中转发项维持不变，终端 C 无法通过新构建的生成树实现和其他终端的通信，因为网桥 B5 根据转发表将目的 MAC 地址除 MAC C 外的所有 MAC 帧通过端口 3 转发出去，这些从端口 3 转发出去的 MAC 帧由于无法被网桥 B3 端口 1(网桥 B3 端口 1 为阻塞端口)接收而丢弃。终端之间通过新构建的生成树实现通信的前提是，在重新构建生成树后，可以让每一个终端发送一个广播帧，让每一个网桥根据新的生成树重新建立转发表；或者让每一个网桥的转发表中的每项转发项因为关联的定时器溢出而被删除，导致新构建的生成树以广播传输方式实现终端之间数据帧的传输。第一个前提是无法实现的，因为生成树结构对终端是透明的，因而终端感觉不到生成树结构的改变。转发项关联定时器的默认值是 300s，减小该值将导致大量终端之间传输的数据帧以广播传输方式进行传输。因此手工配置较小的转发项关联定时器的初值是不可取的，较好的办法是，在生成树结构没有发生变化的情况下，定时器初值采用较合理的值，如默认值。一旦生成树结构发生变化，临时减小定时器初值，加快转发项关联的定时器的溢出速度。这就需要网桥能够做到以下两点：一是能够监测到生成树结构发生改变，并把监测结果通知其他网桥；二是在接收到其他网桥关于生成树结构发生变化的通知后，立即减小转发项关联的定时器的初值。

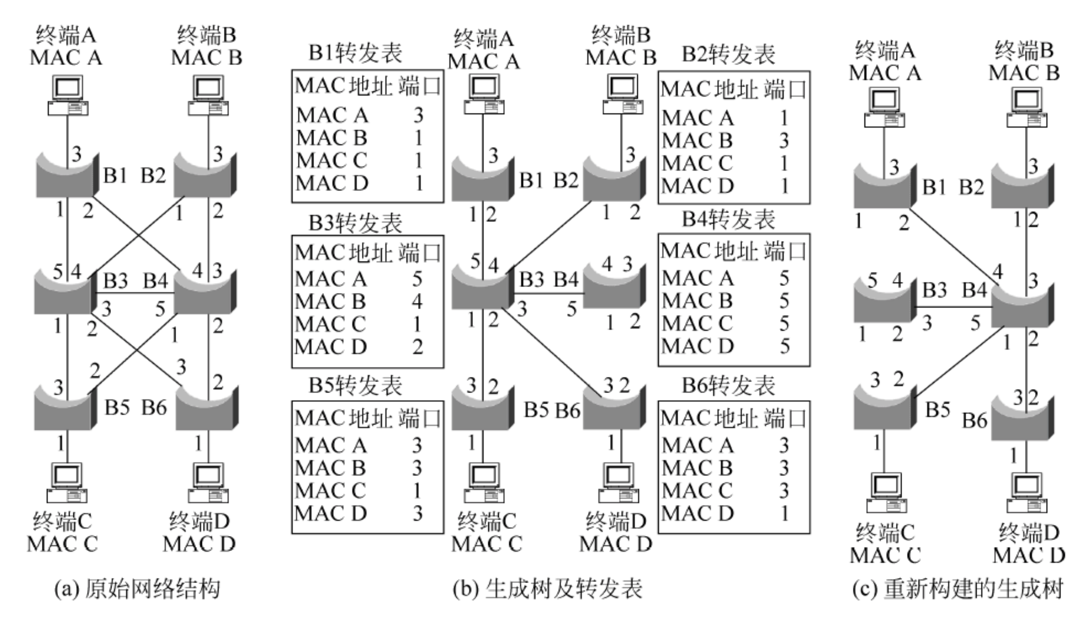


图 3.8 重新构建生成树与转发表刷新

2. 网桥转发表刷新过程

STP 存在两种类型的 BPDU，一种是配置 BPDU，另一种是拓扑改变通知(Topology Change Notification，TCN) BPDU，到目前为止，所讨论的 BPDU 都是配置 BPDU。配置

BPDU 由根网桥定时触发,从根网桥开始向最外围网桥辐射。拓扑改变通知 BPDU 用于通知根网桥生成树结构已经发生变化,从监测到生成树结构发生改变的网桥开始,向根网桥逐跳传输。

某个网桥在检测到下述情况时,确定生成树结构发生变化。

① 从非根网桥转变为根网桥;

② 监测到有端口从其他状态转变为转发状态;

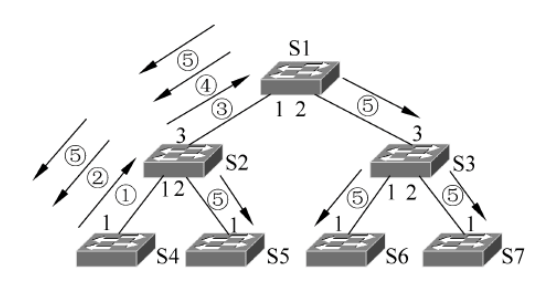
③ 监测到有端口从其他状态转变为阻塞状态。

如果某个非根网桥监测到情况②和情况③,则会一直定时通过根端口发送拓扑改变通知 BPDU,直到接收到拓扑改变应答 BPDU,拓扑改变应答 BPDU 是配置 BPDU,只是置位了其中拓扑改变应答(Topology Change Acknowledgment,TCA)标志位。

如果某个非根网桥通过某个指定端口接收到拓扑改变通知 BPDU,当需要通过该指定端口发送端口 BPDU 时,置位该端口 BPDU 的 TCA 标志位。如果某个发送了拓扑改变通知 BPDU 的网桥通过根端口接收到置位 TCA 标志位的配置 BPDU,表示上游网桥已经接收到拓扑改变通知 BPDU,不再定时发送拓扑改变通知 BPDU。同样,所有通过某个指定端口接收到拓扑改变通知 BPDU 的网桥,一直定时通过根端口发送拓扑改变通知 BPDU,直到收到拓扑改变应答 BPDU(置位 TCA 标志位的配置 BPDU)。

如果某个网桥监测到情况①,或者根网桥通过某个指定端口接收到拓扑改变通知 BPDU,根网桥首先在通过该指定端口发送的第一个端口 BPDU 时同时置位 TCA 标志位和拓扑改变(Topology Change,TC)标志位。通过其他指定端口发送的端口 BPDU 中置位 TC 标志位。根网桥持续 Forward Delay + Max Age 时间定时发送置位 TC 标志位的配置 BPDU,所有通过根端口接收到置位 TC 标志位的配置 BPDU 的网桥,在通过各自指定端口发送端口 BPDU 时,置位端口 BPDU 的 TC 标志位。任何网桥在接收到置位 TC 标志位的配置 BPDU 时间段内,将转发项关联的定时器初值减小为 Forward Delay。

图 3.9 给出了网桥转发表刷新过程。假定图 3.9 中的 S1 是根网桥,S4 监测到拓扑结构发生变化,S4 通过根端口(端口 1)每间隔 Hello Time 时间发送拓扑改变通知 BPDU,S2 通过指定端口(端口 1)接收到 S4 发送的 TCN BPDU 后,在下一次因为受根网桥发送的配置 BPDU 触发,通过端口 1 发送端口 BPDU 时,置位端口 BPDU 中的 TCA 标志位。S4 通过根端口接收到置位 TCA 的配置 BPDU 后,停止 TCN BPDU 的发送。S2 通过端口 1 接收到 TCN BPDU 后,也通过根端口(端口 3)每间隔 Hello Time 时间发送 TCN BPDU,直到通过根端口接收到置位 TCA 标志位的配置 BPDU。S1 通过端口 1 接收到 TCN BPDU 后,在下一次 Hello Time 定时器溢出时,通过端口 1 发送同时置位 TC 和 TCA 标志位的配置 BPDU,通过端口 2 发送置位 TC 标志位的配置 BPDU,在以后由于 Hello Time 定时器溢出通过端口 1 和端口 2 发送配置 BPDU 时,置位 TC 标志位。S1 持续 Forward Delay + Max Age 时间定时通过端口 1 和端口 2 发送置位 TC 标志位的配置 BPDU。所有网桥因为受根网桥置位 TC 标志位的配置 BPDU 触发,通过所有指定端口发送端口 BPDU 时,同样置位端口 BPDU 中的 TC 标志位。这样,置位 TC 标志位的配置 BPDU 从根网桥开始,向外辐射,直到到达最外围网桥。根网桥和所有接收到置位 TC 标志位的配置 BPDU 的网桥将转发项关联的定时器初值改为 Forward Delay 时间。



①：S4发送的拓扑改变通知BPDU

②：S2发送的置位TCA的配置BPDU（即端口BPDU）

③：S2发送的拓扑改变通知BPDU

④：S1发送的置位TCA和TC的配置BPDU

⑤：根网桥定时发送的、其他网桥被触发发送的置位TC的配置BPDU

图 3.9　网桥转发表刷新过程

3.2.6　STP 例题解析

【例 3.1】 网络结构如图 3.10 所示，假定所有网桥的优先级采用默认值，所有端口连接 100Mb/s 链路，整个网络属于默认 VLAN——VLAN 1，求出 STP 构建的生成树，给出每一个端口的状态(D：指定端口，R：根端口，B：阻塞端口)。

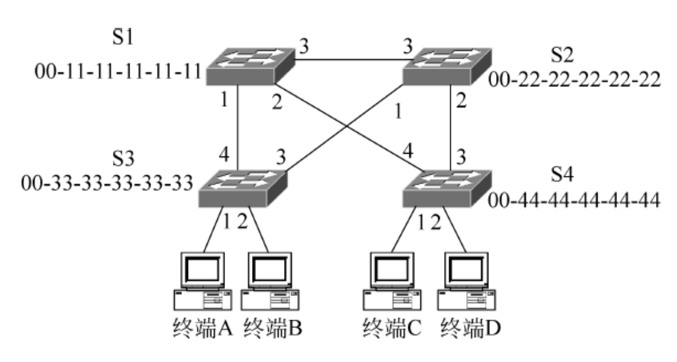


图 3.10　网络结构

【解析】 ① 确定根网桥，由于所有网桥的优先级采用相同的默认值，拥有最小的 MAC 地址的网桥为根网桥，因此，网桥 S1 为根网桥。

② 确定根端口，由于所有链路的路径距离相同，因此，与根网桥之间具有最小跳数的路径即为根路径，连接根路径的端口即为根端口。由于网桥 S2、网桥 S3、网桥 S4 与网桥 S1 直接相连，因此这些网桥连接 S1 的链路的端口：网桥 S2 端口 3、网桥 S3 端口 4、网桥 S4 端口 4，为根端口。

③ 确定指定端口，网桥 S1 的所有端口为指定端口(除非 S1 端口之间直接用链路互连)。网桥 S3、网桥 S4 连接终端的端口为指定端口，因为这些端口接收不到配置 BPDU。网桥 S2、网桥 S3 和网桥 S4 的其他端口通过比较链路两端端口的端口 BPDU 确定指定端口。如比较网桥 S4 端口 3 和网桥 S2 端口 2 的端口 BPDU 时发现，两者的根网桥相同；两者的根路径距离相同(都是一跳，值为 19)；但两者的发送网桥标识符不同，因此较小网桥标识符的网桥 S2 为指定网桥，网桥 S2 端口 2 为指定端口，从而导出网桥 S3 端口 3 为阻塞端

口。其他端口的状态依照此方法确定。

完成上述操作后,最终形成图 3.11(a)所示的端口状态。删除一端为阻塞端口的链路,构成图 3.11(b)所示的生成树。

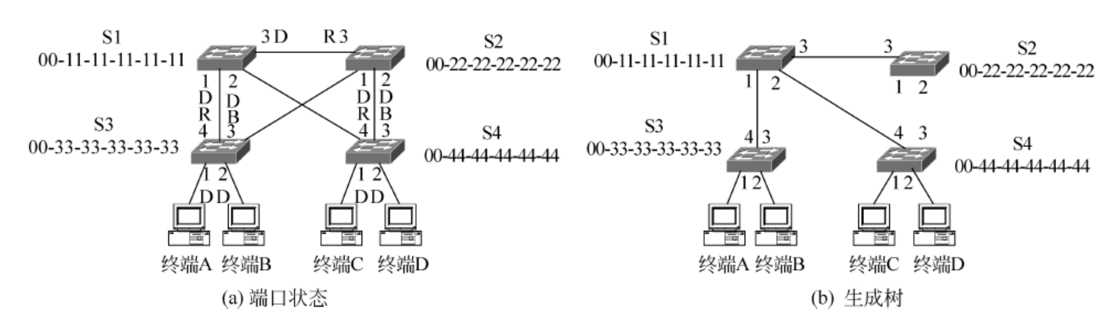


图 3.11 端口状态和生成树

【例 3.2】 网络结构如图 3.10 所示,所有端口连接 100Mb/s 链路,终端 A 和终端 C 属于 VLAN 2,终端 B 和终端 D 属于 VLAN 3,要求不同 VLAN 的流量尽量利用所有链路的带宽,给出需要的配置,求出每一个 VLAN 对应的生成树。

【解析】 由于 VLAN 相当于独立的以太网,因此,不同的 VLAN 对应不同的生成树,生成树是基于 VLAN 的。为了充分利用各条链路的带宽,要求 VLAN 2 对应的生成树的根网桥为 S1,VLAN 3 对应的生成树的根网桥为 S2。因此,在构建基于 VLAN 3 的生成树时,将网桥 S2 的优先级配置为 600(小于默认值)。由此产生如图 3.12 所示的 VLAN 2 和 VLAN 3 对应的生成树。

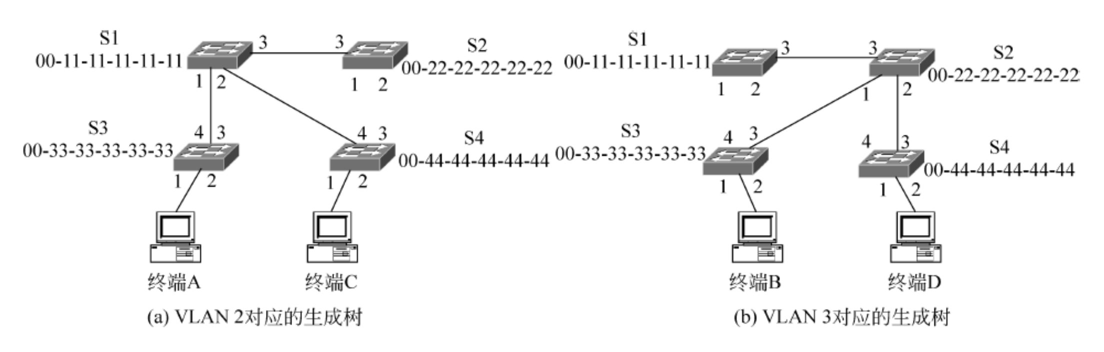


图 3.12 VLAN 2 和 VLAN 3 对应的生成树

3.3 快速生成树协议

3.3.1 STP 的缺陷

STP 收敛前,可能存在短暂的连通和环路问题,为了消除收敛前可能发生的短暂的环路问题,STP 对非阻塞端口(根网桥端口、根端口和指定端口)增加了侦听和学习这两个过渡状态,使得从确定为非阻塞端口到端口开始转发数据帧之间的时间间隔为 2×Forward Delay 时间,其中 Forward Delay 是配置 BPDU 从根网桥辐射到最外围网桥所需要的时间。

以此保证只有在 STP 已经收敛的情况下，某个端口才有可能处于转发状态，消除了短暂环路问题。但 STP 消除短暂环路问题的方法不仅没有消除短暂连通问题，在网络拓扑结构发生变化时，反而有可能使网络 Max Age＋2×Forward Delay 时间存在连通问题。为了在消除短暂环路问题的前提下，有效减少因为网络拓扑结构发生变化而导致的网络存在连通问题的时间，提出了快速生成树协议（Rapid Spanning Tree Protocol，RSTP）。

3.3.2 端口角色和端口状态

1. 端口角色

RSTP 将端口角色分为根端口（Root Port）、指定端口（Designated Port）、替换端口（Alternate Port）、备份端口（Backup Port）和边缘端口（Edge Port），各种端口角色的含义如图 3.13 所示。根端口（图 3.13 中用 R 表示）和指定端口（图 3.13 中用 D 表示）的名称与含义与 STP 相同。替换端口（图 3.13 中用 A 表示）的含义等同于 STP 的阻塞端口。如果同一网桥有多个端口连接到共享网段上，其中端口标识符最小的端口成为指定端口，其他端口则为备份端口（图 3.13 中用 B 表示）。根网桥的所有端口或是指定端口，或是备份端口，假如直接用链路互连根网桥的两个端口，其中一个端口标识符较小的端口为指定端口，另一个为备份端口。备份端口是对指定端口的备份，如果指定端口发生故障，可以用备份端口取代指定端口，这个过程瞬时完成，因此可以大大减少网络存在连通问题的时间。边缘端口（图 3.13 中用 E 表示）是指网桥直接连接终端的端口，这些端口不会构成数据帧传输环路，因此，不需要参与 STP 构建生成树过程。边缘端口通过人工配置确定，如果某个边缘端口接收到 BPDU，意味着网络结构发生问题，应立即关闭该端口，以免产生数据帧传输环路。

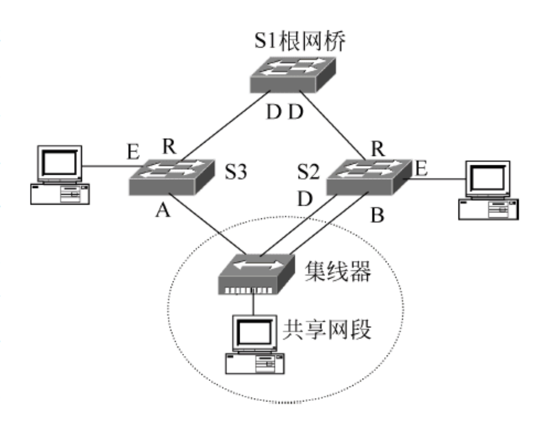


图 3.13 端口类型示意图

2. 端口状态

RSTP 将端口状态简化为三种：丢弃状态（Discarding）、学习状态（Learning）和转发状态（Forwarding），根端口和指定端口可以处于这三种状态的任何一种，处于丢弃状态的根端口和指定端口只允许发送、接收 BPDU。处于学习状态的根端口和指定端口允许发送、接收 BPDU，且允许学习数据帧的源 MAC 地址，但不允许转发数据帧。处于转发状态的根端口和指定端口允许输入、输出数据帧。替换端口和备份端口的稳定状态是丢弃状态，处于丢弃状态的替换端口和备份端口只允许接收 BPDU。边缘端口开通后，立即处于转发状态。

端口角色确定过程，就是 RSTP 构建生成树过程。该过程与 STP 构建生成树过程基本相同。RSTP 与 STP 最大改进在于端口状态迁移过程。STP 中，某个端口被确定为根端口和指定端口后，必须经过 2×Forward Delay 时间才能进入转发状态，RSTP 允许根端口和指定端口快速完成从丢弃状态到转发状态的迁移。

3.3.3 端口状态快速迁移过程

1. BPDU 标志字段

RSTP RPDU 标志字段(图 3.14)增加了 BPDU 发送端口角色和状态标志位,如果 BPDU 发送端口是指定端口且处于丢弃状态,则端口角色标志位值为 11,端口状态标志位 Learning 和 Forwarding 的值为 0。标志位 Proposal 和 Agreement 用于完成指定端口从丢弃状态到转发状态的快速迁移。

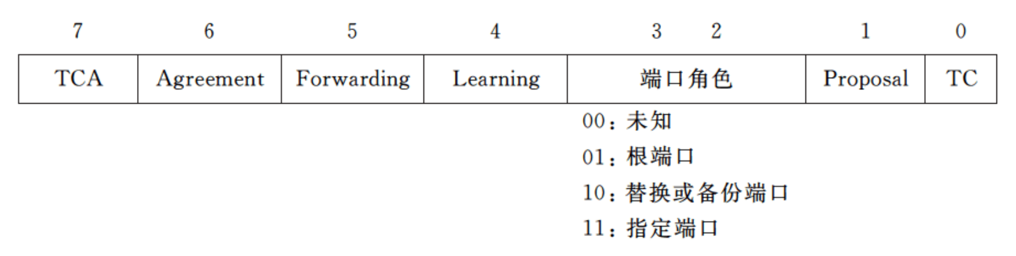


图 3.14　BPDU 标志字段

2. 端口 BPDU 触发机制

STP 中,只有根网桥每间隔 Hello Time 时间通过所有指定端口发送端口 BPDU,非根网桥只有通过根端口接收到 BPDU 时,才通过所有指定端口发送端口 BPDU。因此,BPDU 都是从根网桥辐射到最外围网桥。RSTP 中,每一个网桥每间隔 Hello Time 时间通过所有指定端口发送端口 BPDU,因此,网桥最佳 BPDU 和端口最佳 BPDU 的溢出定时器初值定义为 3×Hello Time,而不是 Max Age 时间,表示如果某个非根网桥持续 3×Hello Time 时间没有通过根端口接收到网桥最佳 BPDU,那么该网桥将回到初始化状态,将自己作为根网桥,开始新的生成树构建过程。同样,如果某个替换或备份端口持续 3×Hello Time 时间没有接收到使其成为替换或备份端口的端口最佳 BPDU,那么该端口将成为指定端口,并定时发送端口 BPDU。

需要强调的是,如果某个端口接收到的 BPDU 次于该端口的端口 BPDU,那么该端口将成为指定端口,并立即发送该端口的端口 BPDU,链路另一端端口接收到该 BPDU 后将转变成替换或备份端口。

3. 指定端口状态快速迁移过程

指定端口快速迁移机制作用于点对点链路,如果某个端口是指定端口,且端口状态处于丢弃或学习状态,该端口发送一个置位 Proposal 标志位的端口 BPDU(称为 Proposal BPDU)。

如果接收到 Proposal BPDU 的链路另一端端口是下述端口角色之一时:

- 根端口;
- 替换端口;
- 备份端口。

该端口所在交换机将所有其他指定端口的状态设置为丢弃状态,然后,向 Proposal BPDU 的发送端口回送一个置位 Agreement 标志位的 BPDU(称为 Agreement BPDU)。发送 Proposal BPDU 的端口一旦接收到链路另一端端口发送的 Agreement BPDU,就将端口

状态直接转变为转发状态，而 STP 将指定端口状态从丢弃状态迁移到转发状态需要 2×Forward Delay 时间。该过程从根网桥指定端口开始，一直辐射到最外围网桥，图 3.15 给出了生成树指定端口状态的快速迁移过程。

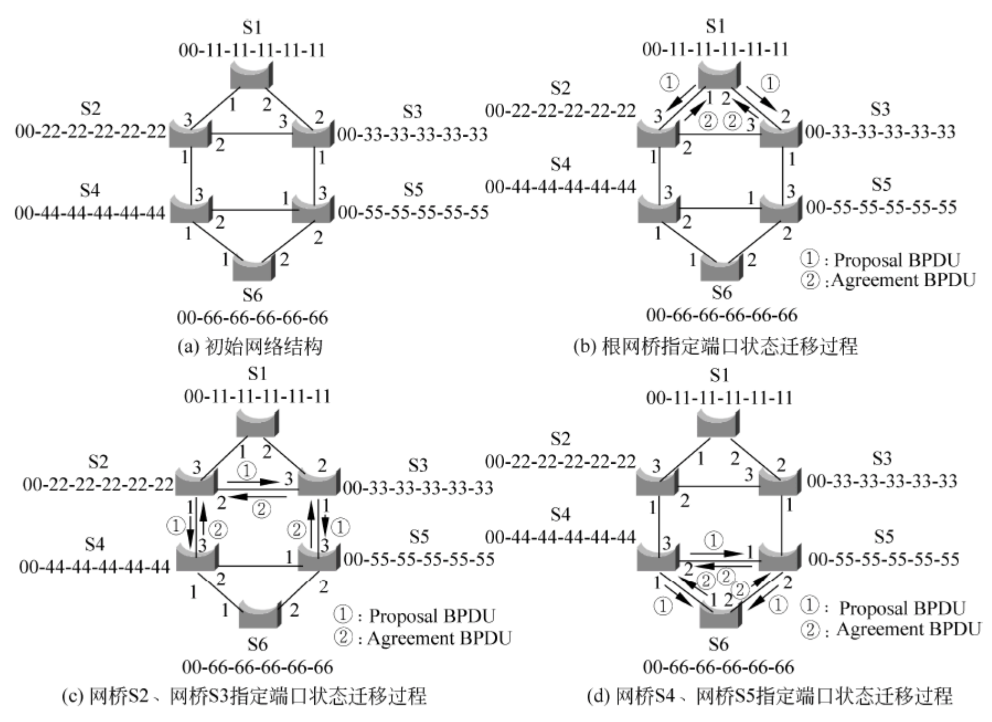


图 3.15 指定端口状态快速迁移过程

当网桥 S1 确定自己为根网桥时，由于网桥 S1 不存在直接用链路互连的端口，所有端口成为指定端口，这些指定端口的初始状态为丢弃状态，因此，网桥 S1 通过这些指定端口发送 Proposal BPDU。网桥 S2 端口 3 和网桥 S3 端口 2 接收到网桥 S1 发送的 Proposal BPDU，由于网桥 S2 端口 3 是根端口，网桥 S2 将指定端口（端口 1 和端口 2）的状态设置为丢弃状态，向网桥 S1 端口 1 发送 Agreement BPDU，网桥 S1 通过端口 1 接收到网桥 S2 端口 3 发送的 Agreement BPDU 后，将端口 1 的状态设置为转发状态。同样，由于网桥 S3 端口 2 是根端口，网桥 S3 将指定端口（端口 1）的状态设置为丢弃状态，向网桥 S1 端口 2 发送 Agreement BPDU，网桥 S1 通过端口 2 接收到网桥 S3 端口 2 发送的 Agreement BPDU 后，将端口 2 的状态设置为转发状态。此时，根网桥的所有指定端口状态已经转换成转发状态，整个过程如图 3.15(a)所示。之所以将网桥 S2 和网桥 S3 除根端口以外的所有其他指定端口的状态设置为丢弃状态，是为了避免因为将根网桥指定端口状态设置为转发状态而产生的数据帧传输环路。

网桥 S2 端口 1 和端口 2 为指定端口，且端口状态为丢弃状态，网桥 S2 通过端口 1 和端口 2 发送 Proposal BPDU，网桥 S2 端口 1 发送的 Proposal BPDU 被网桥 S4 端口 3 接收到，由于 S4 端口 3 是根端口，网桥 S4 将指定端口（端口 1 和端口 2）的状态设置为丢弃状态，向网桥 S2 端口 1 发送 Agreement BPDU，网桥 S2 通过端口 1 接收到网桥 S4 端口 3 发送的 Agreement BPDU 后，将端口 1 的状态设置为转发状态。网桥 S2 通过端口 2 发送的 Proposal BPDU 被网桥 S3 端口 3 接收到，由于网桥 S3 端口 3 是替换端口，网桥 S3 将指定端口（端口 1）的状态设置为丢弃状态，向网桥 S2 端口 2 发送 Agreement BPDU，网桥 S2 通

过端口 2 接收到网桥 S3 端口 3 发送的 Agreement BPDU 后，将端口 2 的状态设置为转发状态。此时，网桥 S2 所有指定端口的状态已经转换成转发状态，整个过程如图 3.15(b)所示。其他网桥依照此方法操作，使得所有网桥的指定端口状态全部转变为转发状态。

4. 根端口状态快速迁移过程

当根端口接收到 BPDU，如果该 BPDU 的标志位表明发送该 BPDU 的端口是指定端口且端口状态为转发状态，根端口状态立即转变为转发状态。如果网桥通过某个非根端口接收到 BPDU，且该 BPDU 成为网桥新的网桥最佳 BPDU，接收该 BPDU 的端口成为新的根端口，只要原来的根端口状态为丢弃状态，且该 BPDU 的标志位表明发送该 BPDU 的端口是指定端口且端口状态为转发状态，新的根端口状态立即转变为转发状态。

3.3.4 网桥转发表刷新机制

当网桥某个非边缘端口的状态迁移到转发状态，表明生成树结构发生改变，需要对网桥的转发表进行刷新操作。监测到生成树结构发生改变的网桥清空转发表中通过非边缘端口学习到的 MAC 地址，同时，通过所有处于转发状态的指定端口和根端口持续 2×Hello Time 时间发送 TC 标志位置位的 BPDU(称为 TC BPDU)。

一旦其他网桥通过处于转发状态的指定端口和根端口接收到 TC BPDU，在转发表中清除除接收该 TC BPDU 端口以外的其他所有非边缘端口学习到的 MAC 地址，通过处于转发状态的指定端口和根端口持续 2×Hello Time 时间发送 TC BPDU。

假定图 3.16 中互连网桥 S4 和 S6 的链路发生故障，网桥 S6 的端口 2 由替换端口变为根端口，并快速完成丢弃状态至转发状态的状态迁移过程，由于网桥 S6 监测到有端口从其他状态转变为转发状态，启动 TC BPDU 泛洪过程，首先网桥 S6 通过处于转发状态的端口(端口 2)发送 TC BPDU，由于网桥 S5 接收到 TC BPDU 的端口是处于转发状态的指定端口，清除转发表中通过除端口 2 以外其他所有端口学习到的 MAC 地址，然后，通过其他所有处于转发状态的指定端口和根端口发送 TC BPDU。由于网桥 S5 除端口 2 外，只有端口 3 是处于转发状态的根端口，通过端口 3 发送 TC BPDU。其他网桥依照此方法泛洪 TC BPDU。网桥 S2 通过端口 2 发送的 TC BPDU 到达网桥 S3 的端口 3，由于网桥 S3 的端口 3 是替换端口，网桥 S3 只是简单丢弃该 TC BPDU，没有后续处理过程。只处理通过处于转发状态的指定端口和根端口接收到的 TC BPDU，只从除接收 TC BPDU 端口以外的其他所有处于转发状态的指定端口和根端口泛洪 TC BPDU，保证 TC BPDU 只沿着生成树传播。

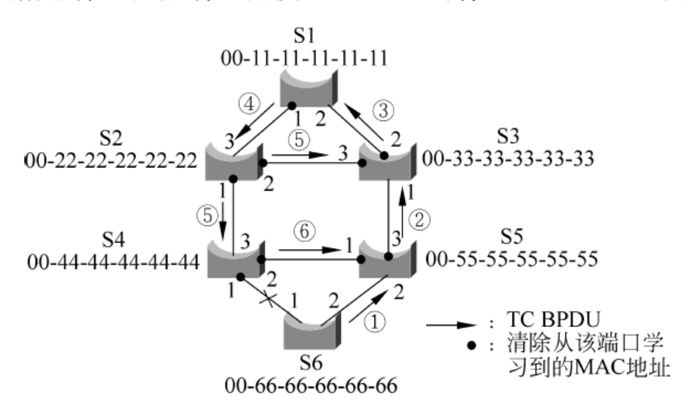


图 3.16 TC BPDU 泛洪过程

3.3.5 RSTP 例题解析

【例 3.3】 网络结构如图 3.17 所示，网桥 S1、网桥 S2 和网桥 S3 的优先级分别为 100、200 和 300，网桥之间链路的传输速率全部为 100Mb/s，如果在完成生成树构建后，拔掉网桥 S1 和网桥 S3 之间的链路，给出 RSTP 下网桥 S3 端口角色和状态的变化过程。

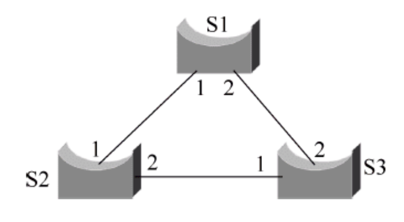


图 3.17 网络结构

【解析】 一旦拔掉网桥 S1 和网桥 S3 之间的链路，网桥 S3 将监测到端口 2 处于链路断开状态，需要重新寻找新的根端口，由于网桥 S3 中存在替换端口，表明存在其他通往根网桥的路径，网桥 S3 在所有替换端口中选择保存的端口最佳 BPDU 最优的替换端口作为根端口，并立即将其状态转变为转发状态。这里，网桥 S3 只有端口 1 是替换端口，立即将端口 1 作为根端口，并使其处于转发状态，因此，RSTP 下，瞬时完成端口角色和状态转换。

【例 3.4】 如图 3.17 所示网络结构，在完成生成树构建后，拔掉网桥 S1 和网桥 S2 之间的链路，分别给出 STP 和 RSTP 下网桥 S2 端口角色和状态的变化过程。

【解析】 ① STP 下，网桥 S2 监测到端口 1 处于链路断开状态，且不存在其他通往根网桥的路径，因此，将自己作为根网桥，并通过所有连接链路的端口发送以网桥 S2 为根网桥的端口 BPDU，这里网桥 S2 只通过端口 2 发送端口 BPDU{S2，0，S2，2}（根网桥是 S2，根路径距离=0，发送网桥标识符是 S2，发送端口标识符是端口 2），由于网桥 S3 端口 1 为替换端口，且端口保存的端口最佳 BPDU 是{S1，19，S2，2}（根网桥是 S1，根路径距离=19，发送网桥标识符是 S2，发送端口标识符是端口 2），当网桥 S3 通过端口 1 接收到 BPDU{S2，0，S2，2}，由于{S2，0，S2，2}次于{S1，19，S2，1}，因此，不会对网桥 S3 产生影响。直到经过 Max Age 时间，{S1，19，S2，2}关联的定时器溢出，网桥 S3 端口 1 的端口 BPDU{S1，19，S3，1}（根网桥是 S1，根路径距离=19，发送网桥标识符是 S3，发送端口标识符是端口 1）成为端口最佳 BPDU，端口 1 成为指定端口，并在通过端口 2 接收到根网桥 S1 发送的 BPDU 的情况下，通过端口 1 发送端口 BPDU{S1，19，S3，1}，使得网桥 S2 将端口 2 确定为根端口，并经过 2×Forward Delay 时间使根端口（端口 2）进入转发状态。

RSTP 下，一旦确定互连网桥 S2 和 S3 的链路为点对点链路，当网桥 S3 通过端口 1 接收到 BPDU{S2，0，S2，2}，发现{S2，0，S2，2}次于端口 1 的端口 BPDU{S1，19，S3，1}，网桥 S3 端口 1 立即成为指定端口，将状态设置为丢弃状态，并通过端口 1 发送置位 Proposal 标志位的端口 BPDU{S1，19，S3，1}，当网桥 S2 通过端口 2 接收到 BPDU{S1，19，S3，1}，将该 BPDU 作为网桥最佳 BPDU，将端口 2 作为根端口，回送 Agreement BPDU，使网桥 S3 端口 1 立即进入转发状态。一旦网桥 S3 端口 1 进入转发状态，通过向网桥 S2 端口 2 发送表明端口角色是指定端口、端口状态是转发状态的 BPDU，使网桥 S2 端口 2 进入转发状态。

3.4 多生成树协议

3.4.1 MSTP 的必要性

STP 和 RSTP 是单生成树协议，基于整个物理以太网构建单个生成树，这样做，一是无

法做到将属于不同 VLAN 的流量均衡分布到多条不同的链路上；二是不同 VLAN 之间的传输路径可能经过不同的链路，一旦基于整个物理以太网构建单个生成树，在保证一些 VLAN 的连通性的情况下，可能导致其他一些 VLAN 无法保证连通性。虽然，一些设备厂家，如 Cisco，将 STP 和 RSTP 扩展为基于 VLAN 构建生成树，但由于不同 VLAN 构建生成树的操作相互独立，导致经过共享链路的 BPDU 流量剧增，影响网络性能。因此，需要一种既是基于 VLAN 构建生成树，又尽可能降低生成树构建操作开销的生成树协议，这就是多生成树协议(Multiple Spanning Tree Protocol，MSTP)产生的原因。

3.4.2 MSTP 基本思想

1. 基本概念

实施 MSTP 的网络结构如图 3.18 所示，一是 MSTP 将网络分成若干域，每一个域由若干交换机和互连这些交换机的链路组成，每一台交换机只能属于单个域。MSTP 将生成树构建过程分成了几个阶段，首先，构建公共生成树(Common Spanning Tree，CST)，在构建 CST 过程中，域等同于一个结点，即等同于 STP 和 RSTP 构建生成树过程中的一台交换机。对于 CST，域有着统一的外特性。对于每一个域，一个或若干个 VLAN 可以映射到同一棵生成树，但每一个 VLAN 只能映射到单棵生成树。域内 MSTP 构建的基于一个或一组 VLAN 的生成树称为多生成树实例(Multiple Spanning Tree Instance，MSTI)。其中，内部生成树(Internal Spanning Tree，IST)是一种特殊的多生成树实例其特殊之处在于：一是无论是否建立 VLAN 与生成树之间映射，该生成树都会建立；二是所有没有和特定生成树建立映射的 VLAN 都映射到该生成树。每一个域内的 IST 和 CST 结合，建立保证所有交换机之间连通性的公共内部生成树(Common and Internal Spanning Tree，CIST)，所有 MSTP BPDU 沿着 CIST 传输，即 CIST 中处于学习和转发状态的端口接收和发送 MSTP BPDU，处于丢弃状态的端口只接收 MSTP BPDU。图 3.19 就是根据图 3.18 所示网络结构构建的 CIST。对于域 2，构建了三棵生成树：一是 IST，二是 VLAN 2 映射的生成树，三是 VLAN 3 映射的生成树。这三棵生成树可以有不同的根交换机，域内有着不同的传输路径。每一个域中 VLAN 2 映射的生成树如图 3.20(a)所示，VLAN 3 映射的生成树如图 3.20(b)所示。CIST 的根交换机称为总根，它是全网络中优先级最高的交换机。每一个域内距离总根最近的交换机称为该域的主交换机。由于每一个域可以基于 VLAN 构建生成树，因此，不同 VLAN 可以映射到不同的生成树，每一个多生成树实例在域内有着自己的根交换机，该交换机称为该多生成树实例对应的域根。如果所有互连交换机的链路的带宽相同的话，可以看出交换机 S1 是 CIST 的总根。交换机 S3 是域 1 中 VLAN 2 映射的多生成树实例的域根，交换机 S5 是域 2 中 VLAN 2 映射的多生成树实例的域根，交换机 S9 是域 3 中 VLAN 2 映射的多生成树实例的域根。

2. 端口角色和状态

域内的每一个多生成树实例存在独立的域根，同样，域内的交换机对应每一个多生成树实例都存在根端口、指定端口、替换端口和备份端口，这些端口角色的功能及状态与 STP 和 RSTP 完全相同。对于必须构建的 CIST，增加了以下端口角色。

- 总根。网络结构中优先级最高的交换机。

• 主网桥。每一个域中距离总根最近的交换机，它同时是 IST 的域根。对于总根所在的域，主网桥和总根是同一个交换机。

• 主端口。主网桥连接 CST 的端口，位于该域通往总根的最短路径上。由于构建 CST 时将域等同于一个结点，因此，主端口是该域在 CST 中的根端口。

• 域边界端口。位于域的边缘，用于和其他域相连，主端口是域边界端口，但所有的域边界端口中只有一个域边界端口是主端口。值得强调的是，由于每一个多生成树实例只具有域内意义，不同域内，同一 VLAN 映射的多生成树实例也是相互独立的，如图 3.20(a)所示域 2 内 VLAN 2 映射的多生成树实例和域 3 内 VLAN 2 映射的多生成树实例。因此，构建多生成树实例时不会涉及域边界端口的角色。为了保证不同域内同一 VLAN 映射的多生成树实例之间的连通性，所有域边界端口在所有多生成树实例中的角色必须与这些域边界端口在 CIST 中的角色保持一致。

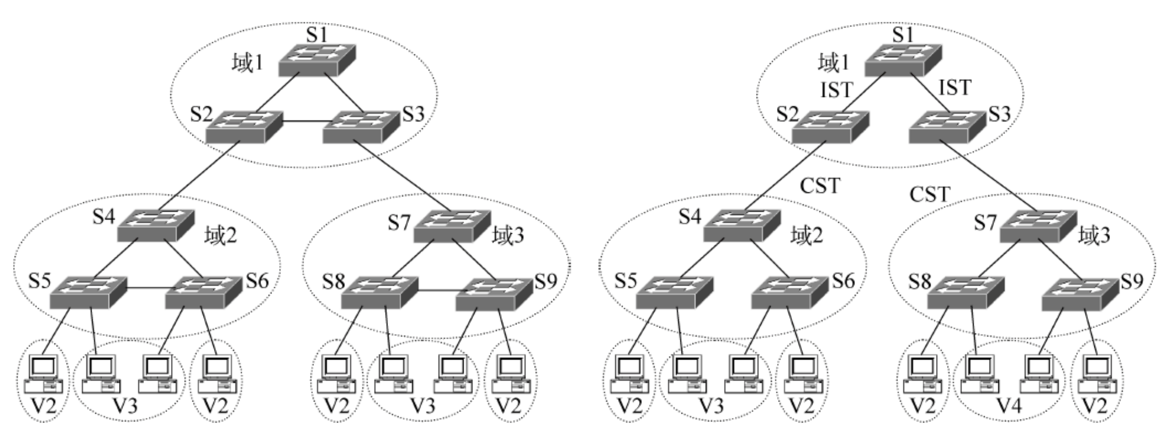
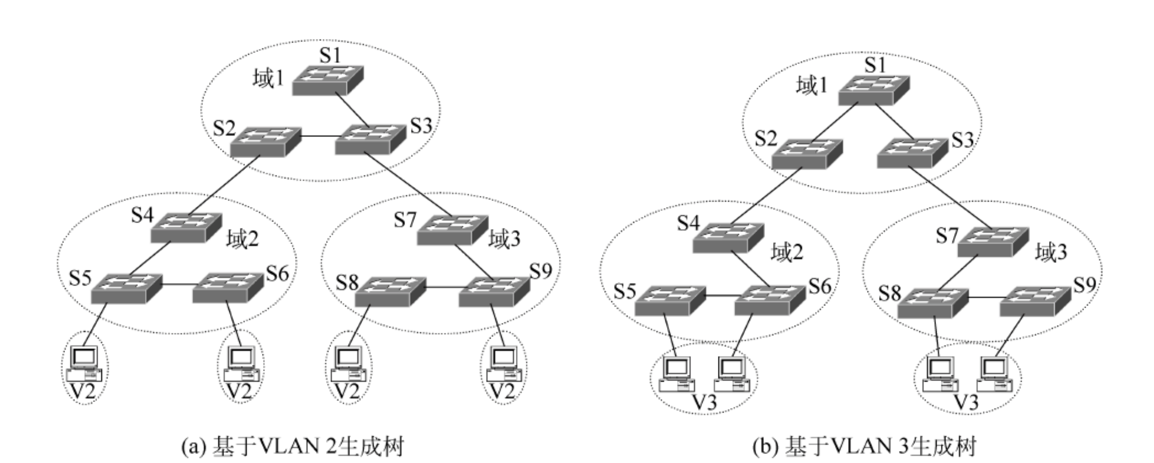


图 3.18　实施 MSTP 的网络结构　　　　图 3.19　CIST



(a) 基于VLAN 2生成树　　　　(b) 基于VLAN 3生成树

图 3.20　基于 VLAN 生成树

3.4.3　MSTP 工作过程

1. MSTP BPDU 格式

MSTP BPDU 包含两部分信息：一是用于构建 CIST 的信息，称为域间信息；二是用于

构建 MSTI 的信息，称为域内信息。由于每一个域可以同时构建多个 MSTI，因此，MSTP BPDU 可能包含多组域内信息。MSTP BPDU 包含的两部分信息如图 3.21 所示。

CIST 标志
CIST 根标识符
CIST 外部路径距离
CIST 域根标识符
CIST 发送端口标识符
域配置信息
CIST 内部路径距离
CIST 发送网桥标识符

(a) 域间信息

MSTI 标志
MSTI 域根标识符
MSTI 内部路径距离
MSTI 发送网桥标识符
MSTI 发送端口标识符
MSTI 剩余跳数
MSTI 标识符

(b) 域内信息

图 3.21 MSTP BPDU 格式

2. CIST 算法

1) 总根

总根是网络中优先级最高(网桥标识符值最小)的交换机。和 RSTP 一样，初始时，所有网桥将自己作为总根，向外发送 CIST 根标识符为自身标识符的 MSTP BPDU，经过几轮交换 MSTP BPDU，最终确定网络中网桥标识符值最小的交换机为总根，其网桥标识符作为 CIST 根标识符。

2) 域根

每一个域根是该域距离总根最近的交换机。当某个交换机接收到 MSTP BPDU，且该 MSTP BPDU 发送交换机所在的域与接收该 MSTP BPDU 的交换机不在同一个域(通过域边界端口接收到 MSTP BPDU)时，将接收该 MSTP BPDU 的端口的端口路径距离累加到该 MSTP BPDU 中的 CIST 外部路径距离字段。然后，在所有通过域边界端口接收到的 MSTP BPDU 中求出最佳 BPDU。求出最佳 MSTP BPDU 的过程是依次比较这些 MSTP BPDU 中的 CIST 根标识符和 CIST 外部路径距离，具有较小字段值的 MSTP BPDU 为最佳 BPDU，将接收最佳 MSTP BPDU 的域边界端口设置为主端口(域根端口)。以最佳 MSTP BPDU 中的 CIST 根标识符、CIST 外部路径距离作为该交换机以后发送的 MSTP BPDU 中的 CIST 根标识符和 CIST 外部路径距离，将该交换机的网桥标识符作为 CIST 域根标识符，设置 CIST 内部路径距离初值 0。如果某个确定为域根的交换机接收到同一域内交换机发送的 MSTP BPDU，则将接收该 MSTP BPDU 的端口的端口路径距离累加到 CIST 内部路径距离字段，将交换机根据最佳 BPDU 得出的 CIST 根标识符、CIST 外部路径距离和自身网桥标识符与该 MSTP BPDU 中的 CIST 根标识符、CIST 外部路径距离和 CIST 域根标识符依次比较，一旦发现该 MSTP BPDU 更优，将域根端口设置为替换端口，将该 MSTP BPDU 作为最佳 BPDU，用该 MSTP BPDU 中的 CIST 根标识符、CIST 外部路径距离、CIST 域根标识符和 CIST 内部路径距离作为自己以后发送的 MSTP BPDU 中相应字段值。域内所有其他交换机从接收到的来自同一域内交换机的 MSTP BPDU 中求出最佳 MSTP BPDU，以最佳 MSTP BPDU 中的 CIST 根标识符、CIST 外部路径距离和 CIST 域根标识符作为自己以后发送的 MSTP BPDU 中的相应字段值。最终，使得所有域内交换机保持的 CIST 根标识符、CIST 外部路径距离和 CIST 域根标识符相同。

3）构建 IST

构建 IST 的算法与 RSTP 构建生成树的算法完全相同，交换机一旦通过非域边界端口接收到 MSTP BPDU，即将接收该 MSTP BPDU 的端口的端口路径距离累加到 CIST 内部路径距离字段，然后在所有通过非域边界端口接收到的 MSTP BPDU 中求出 IST 最佳 BPDU，通过依次比较所有接收到的 MSTP BPDU 中的 CIST 内部路径距离、CIST 发送网桥标识符和 CIST 发送端口标识符求出 IST 最佳 BPDU，具有较小字段值的 MSTP BPDU 为 IST 最佳 BPDU，交换机将接收 IST 最佳 BPDU 的端口作为根端口，根据 IST 最佳 BPDU 得出的 CIST 内部路径距离作为以后发送的 MSTP BPDU 中的 CIST 内部路径距离字段值。每一端口根据 IST 最佳 BPDU 生成端口 BPDU，如果某个非根端口接收到 MSTP BPDU，且该 MSTP BPDU 优于该端口的端口 BPDU，将端口设置为替换端口，否则，将端口设置为指定端口。

值得强调的是，域内交换机确定域根和 IST 的过程是同步进行的，只是通过域边界端口接收到 MSTP BPDU 时，需要将该端口的端口路径距离累加到 MSTP BPDU 中的 CIST 外部路径距离字段，通过非域边界端口接收到 MSTP BPDU 时，需要将该端口的端口路径距离累加到 MSTP BPDU 中的 CIST 内部路径距离字段。通过依次比较所有接收到的 MSTP BPDU 中的 CIST 根标识符、CIST 外部路径距离、CIST 域根标识符、CIST 内部路径距离、CIST 发送网桥标识符和 CIST 发送端口标识符求出 CIST 最佳 BPDU，具有较小字段值的 MSTP BPDU 为 CIST 最佳 BPDU。接收 CIST 最佳 BPDU 的域边界端口为主端口（域根端口），接收 CIST 最佳 BPDU 的非域边界端口为根端口，每一个交换机根据 CIST 最佳 BPDU 生成端口 BPDU，如果某个非根端口接收到 MSTP BPDU，且该 MSTP BPDU 优于该端口的端口 BPDU，将端口设置为替换端口，否则，将端口设置为指定端口。

3. MSTI 算法

由于只在域内构建 MSTI，所以构建 MSTI 的算法与 RSTP 完全一样，只是交换机确定 MSTI 最佳 BPDU 时，依次比较 MSTI 域根标识符、MSTI 内部路径距离、MSTI 发送网桥标识符、MSTI 发送端口标识符。配置时，对每一个不同的 MSTI 分配一个 MSTI 标识符，将一个或一组 VLAN 与某个 MSTI 标识符绑定，交换机可以针对不同的 MSTI 标识符分配不同的优先级。域内所有交换机为特定 MSTI 标识符分配优先级后生成该 MSTI 对应的网桥标识符，域内所有交换机对应该 MSTI 的最小网桥标识符为该 MSTI 的 MSTI 域根标识符。为了实现 MAC 帧转发，将用于交换机之间互连的端口配置为被所有 VLAN 共享的标记端口。完成 MSTI 构建后，当交换机接收到某个 MAC 帧时，首先确定该 MAC 帧所属的 VLAN，然后查找和该 MAC 帧所属 VLAN 绑定的 MSTI，如果接收该 MAC 帧的端口对于该 MSTI 处于转发状态，继续该 MAC 帧转发过程，否则，丢弃该 MAC 帧。确定该 MAC 帧转发端口后，同样需要判断转发端口对于该 MSTI 是否处于转发状态，只有当转发端口对于该 MSTI 处于转发状态时，才能通过转发端口输出该 MAC 帧。

4. MSTP 构建 CIST 实例

1）初始端口和最佳 BPDU

网络结构如图 3.22 所示，假定交换机标识符：S1<S2<S3<S4<S5<S6，所有交换机端口速率为 100Mb/s，CIST 构建过程如下。对于域 1 和域 2，各个交换机初始最佳 MSTP

BPDU 和端口 BPDU 如表 3.3 所示。

表 3.3 各个交换机初始最佳 MSTP BPDU 和端口 BPDU

交换机	BPDU 类型	MSTP BPDU ＜CIST 根标识符,CIST 外部路径距离,CIST 域根标识符,CIST 内部路径距离,CIST 发送网桥标识符,CIST 发送端口标识符,域名＞
S1	最佳 BPDU	＜S1,0,S1,0,S1＞
	端口 1 端口 BPDU	＜S1,0,S1,0,S1,1,域 1＞
	端口 2 端口 BPDU	＜S1,0,S1,0,S1,2,域 1＞
	端口 3 端口 BPDU	＜S1,0,S1,0,S1,3,域 1＞
S2	最佳 BPDU	＜S2,0,S2,0,S2＞
	端口 1 端口 BPDU	＜S2,0,S2,0,S2,1,域 1＞
	端口 2 端口 BPDU	＜S2,0,S2,0,S2,2,域 1＞
S3	最佳 BPDU	＜S3,0,S3,0,S3＞
	端口 1 端口 BPDU	＜S3,0,S3,0,S3,1,域 1＞
	端口 2 端口 BPDU	＜S3,0,S3,0,S3,2,域 1＞
	端口 3 端口 BPDU	＜S3,0,S3,0,S3,3,域 1＞
S4	最佳 BPDU	＜S4,0,S4,0,S4＞
	端口 1 端口 BPDU	＜S4,0,S4,0,S4,1,域 2＞
	端口 2 端口 BPDU	＜S4,0,S4,0,S4,2,域 2＞
	端口 3 端口 BPDU	＜S4,0,S4,0,S4,3,域 2＞
S5	最佳 BPDU	＜S5,0,S5,0,S5＞
	端口 1 端口 BPDU	＜S5,0,S5,0,S5,1,域 2＞
	端口 2 端口 BPDU	＜S5,0,S5,0,S5,2,域 2＞
S6	最佳 BPDU	＜S6,0,S6,0,S6＞
	端口 1 端口 BPDU	＜S6,0,S6,0,S6,1,域 2＞
	端口 2 端口 BPDU	＜S6,0,S6,0,S6,2,域 2＞
	端口 3 端口 BPDU	＜S6,0,S6,0,S6,3,域 2＞

2）域 1 内部生成树构建过程

交换机 S1、交换机 S2 和交换机 S3 定时发送端口 BPDU。交换机 S2 通过端口 1 接收到交换机 S3 端口 3 发送的端口 BPDU＜S3,0,S3,0,S3,3,域 1＞,发现端口 1 的端口 BPDU＜S2,0,S2,0,S2,1,域 1＞优于接收到的 BPDU＜S3,0,S3,19,S3,3,域 1＞(S2＜S3),因此,交换机 S2 维持最佳 BPDU 不变,端口 1 维持端口角色(指定端口)。交换机 S2 通过端口 2 接收到交换机 S1 端口 3 发送的端口 BPDU＜S1,0,S1,0,S1,3,域 1＞,发现端口 2 的端口 BPDU＜S2,0,S2,0,S2,1,域 1＞劣于接收到的 BPDU＜S1,0,S1,19,S1,3,域 1＞(S2＞S1),将接收端口的端口路径距离 19 累加到接收到的 BPDU 中的 CIST 内部路径距离字段,使得接收到的 BPDU 变为＜S1,0,S1,19,S1,3,域 1＞,并因此导致交换机 S2 维持的最佳 BPDU 变为＜S1,0,S1,19,S1＞,根据最佳 BPDU 推导出端口 1 的端口 BPDU 为＜S1,0,S1,19,S2,1,域 1＞。由于交换机 S2 通过端口 2 接收到的 BPDU 推导出最佳 BPDU,因此,将端口 2 的端口角色设置为根端口。

当交换机 S3 通过端口 2 接收到交换机 S1 端口 1 发送的端口 BPDU＜S1,0,S1,0,S1,

1，域 1＞，发现端口 2 的端口 BPDU＜S3，0，S3，0，S3，2，域 1＞劣于接收到的 BPDU＜S1，0，S1，19，S1，1，域 1＞（S3＞S1），将接收端口的端口路径距离 19 累加到接收到的 BPDU 中的 CIST 内部路径距离字段，使得接收到的 BPDU 变为＜S1，0，S1，19，S1，1，域 1＞。并因此导致交换机 S3 维持的最佳 BPDU 变为＜S1，0，S1，19，S1＞，根据最佳 BPDU 推导出端口 1 的端口 BPDU 为＜S1，0，S1，19，S3，1，域 1＞，端口 3 的端口 BPDU 为＜S1，0，S1，19，S3，3，域 1＞。由于交换机 S3 通过端口 2 接收到的 BPDU 推导出最佳 BPDU，因此，将端口 2 的端口角色设置为根端口。交换机 S3 通过端口 3 接收到交换机 S2 端口 1 发送的端口 BPDU＜S1，0，S1，19，S2，1，域 1＞，发现端口 3 的端口 BPDU＜S1，0，S1，19，S3，3，域 1＞劣于接收到的 BPDU＜S1，0，S1，19，S2，1，域 1＞（S1＝＝S1 · AND · 0＝＝0 · AND · S1＝＝S1 · AND · 19＝＝19 · AND · S3＞S2），但由于累加端口 3 端口路径距离后的 CIST 内部路径距离（19＋19＝38）大于交换机 S3 维持的最佳 BPDU 中的 CIST 内部路径距离（19），交换机 S3 维持最佳 BPDU 不变，将端口 3 的端口角色设置为替换端口。

3）CST 和域 2 内部生成树构建过程

域 2 交换机 S4 端口 3 接收到交换机 S3 端口 1 发送的端口 BPDU＜S1，0，S1，19，S3，1，域 1＞，发现该 BPDU 优于端口 3 的端口 BPDU（S1＜S4），且该 BPDU 的发送交换机与自己不在同一个域，将端口 3 的端口路径距离 19 累加到接收到的 BPDU 中的 CIST 外部路径距离字段，使得接收到的 BPDU 变为＜S1，19，S1，19，S3，1，域 1＞，生成最佳 BPDU＜S1，19，S4，0，S3＞，将端口 3 设置为主端口（域根端口），根据最佳 BPDU 推导出端口 1 和端口 2 的端口 BPDU 分别为＜S1，19，S4，0，S4，1，域 2＞和＜S1，19，S4，0，S4，2，域 2＞。

同样，域 2 交换机 S6 端口 2 接收到交换机 S1 端口 2 发送的端口 BPDU＜S1，0，S1，0，S1，2，域 1＞，发现该 BPDU 优于端口 2 的端口 BPDU（S1＜S6），且该 BPDU 的发送交换机与自己不在同一个域，将端口 2 的端口路径距离 19 累加到接收到的 BPDU 中的 CIST 外部路径距离字段，使得接收到的 BPDU 变为＜S1，19，S1，0，S1，2，域 1＞，生成最佳 BPDU＜S1，19，S6，0，S1＞，将端口 2 设置为主端口（域根端口），根据最佳 BPDU 推导出端口 1 和端口 3 的端口 BPDU 分别为＜S1，19，S6，0，S6，1，域 2＞和＜S1，19，S6，0，S6，3，域 2＞。

域 2 交换机 S6 端口 3 接收到交换机 S4 端口 2 发送的端口 BPDU＜S1，19，S4，0，S4，2，域 2＞，发现该 BPDU 优于端口 3 的端口 BPDU（S1＝＝S1 · AND · 19＝＝19 · AND · S4＜S6），将端口 3 的端口路径距离 19 累加到接收到的 BPDU 中的 CIST 内部路径距离字段，使得接收到的 BPDU 变为＜S1，19，S4，19，S4，2，域 2＞，生成最佳 BPDU＜S1，19，S4，19，S4＞，将端口 3 设置为根端口，将端口 2（原来的域根端口）设置为替换端口，根据最佳 BPDU 推导出端口 1 和端口 3 的端口 BPDU 分别为＜S1，19，S4，19，S6，1，域 2＞和＜S1，19，S4，19，S6，3，域 2＞。

交换机 S5 通过端口 1 接收到交换机 S4 端口 1 发送的端口 BPDU＜S1，19，S4，0，S4，1，域 2＞，发现该 BPDU 优于端口 1 的端口 BPDU（S1＜S5），将端口 1 的端口路径距离 19 累加到接收到的 BPDU 中的 CIST 内部路径距离字段，使得接收到的 BPDU 变为＜S1，19，S4，19，S4，1，域 2＞，生成最佳 BPDU＜S1，19，S4，19，S4＞，将端口 1 设置为根端口，根据最佳 BPDU 推导出端口 2 的端口 BPDU 为＜S1，19，S4，19，S5，2，域 2＞。交换机 S5 通过端口 2 接收到交换机 S6 端口 1 发送的端口 BPDU＜S1，19，S4，19，S6，1，域 2＞，发现该 BPDU 劣于端口 2 的端口 BPDU（S1＝＝S1 · AND · 19＝＝19 · AND · S4＝＝S4 · AND · 19＝＝19 · AND · S6＞S5），将端口 2 设置为指定端口。

交换机 S6 通过端口 1 接收到交换机 S5 端口 2 发送的端口 BPDU＜S1，19，S4，19，S5，2，域 2＞，发现该 BPDU 优于端口 1 的端口 BPDU(S1＝＝S1 · AND · 19＝＝19 · AND · S4＝＝S4 · AND · 19＝＝19 · AND · S5＜S6)，将端口 1 设置为替换端口。

完成上述操作后，得出图 3.23 所示的 CIST。值得强调的是，计算 CST 时，域等同于一个结点，因此，到总根的最短距离，其实是到总根所在域的最短距离，虽然域 2 交换机 S6 到总根 S1 的距离最短，但到域 1 的距离与交换机 S4 到域 1 的距离相等，且交换机 S4 的网桥标识符小于交换机 S6 的网桥标识符，因此，交换机 S4 成为域 2 的域根。

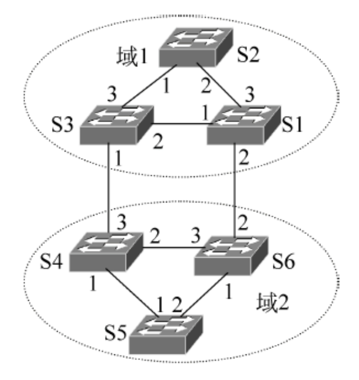


图 3.22 网络结构

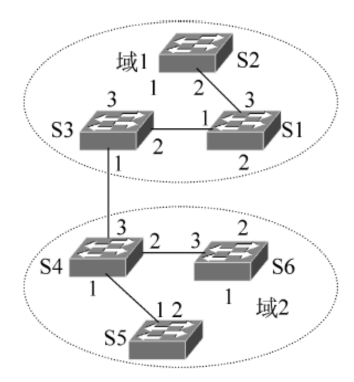


图 3.23 CIST

习题

3.1 简述 STP 确定网桥根端口的步骤。

3.2 简述 STP 确定网桥指定端口和替换端口的步骤。

3.3 STP 中如何计算 Max Age 和 Forward Delay?

3.4 简述拓扑改变通知 BPDU 的作用。

3.5 简述 RSTP 加速生成树收敛的机制。

3.6 STP 和 RSTP 生成配置 BPDU 有什么不同?

3.7 MSTP 构建 MSTI 时有哪些减少 BPDU 流量机制?

3.8 根据图 3.24 标明的各网桥标识符，求出图中网桥所有端口的类型(RP：根端口，DP：指定端口，NDP：非指定端口)和状态(F：转发状态，B：阻塞状态)。

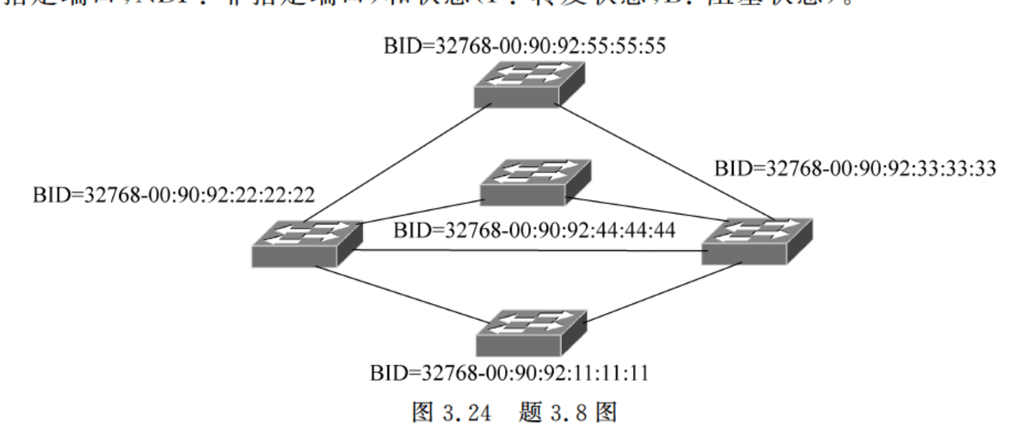


图 3.24 题 3.8 图

3.9 网络结构如图 3.25 所示，整个网络属于一个域。对应图中的每一个 VLAN 生成 MSTI，要求这些 MSTI 尽量均衡每一条链路上的流量。画出所有 MSTI。

3.10 网络结构如图 3.26 所示，假定网桥标识符 S1＜S2＜S3＜S4＜S5，所有链路的带宽相同，画出对应的 CIST。如果将两个域归并为一个域，画出对应的 CIST，并解释造成这两个 CIST 不同的原因。

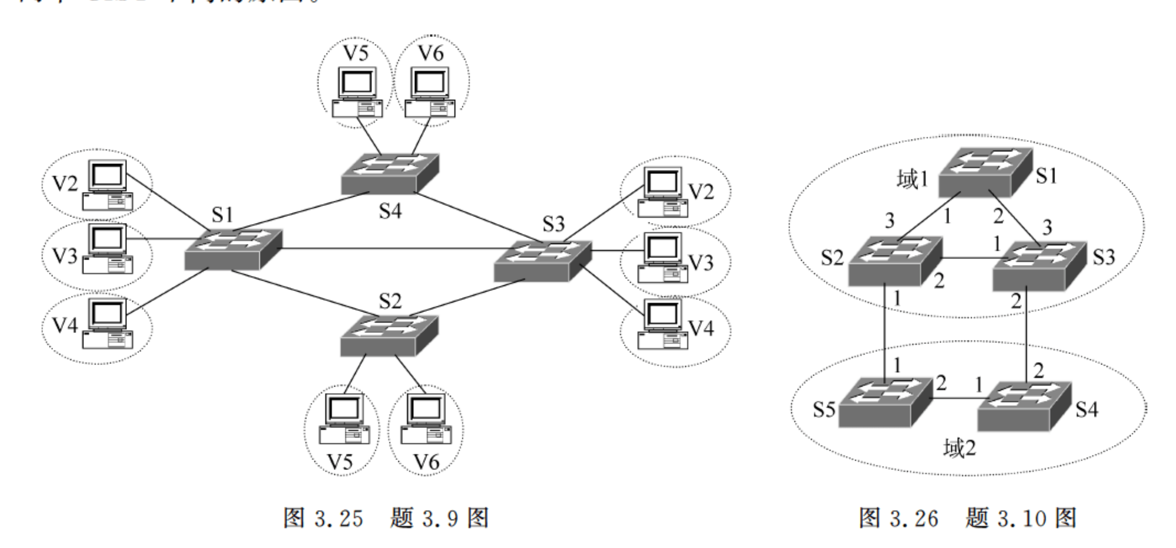


图 3.25 题 3.9 图　　　　图 3.26 题 3.10 图

3.11 网络结构如图 3.27 所示，假定网桥标识符 S1＜S2＜S3＜S4＜S5＜S6＜S7＜S8＜S9＜S10＜S11＜S12＜S13＜S14，所有链路的带宽相同，画出对应的 CIST。

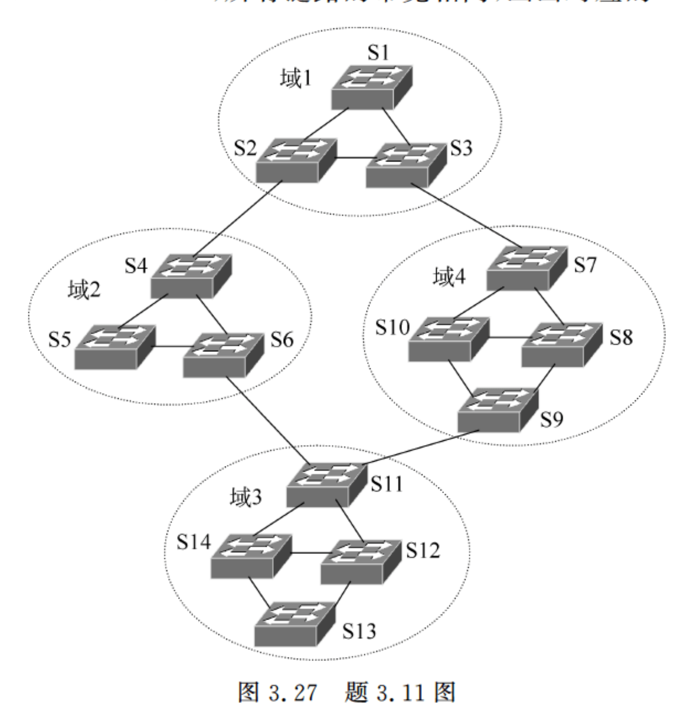


图 3.27 题 3.11 图

第4章 以太网链路聚合

以太网链路聚合技术使得一组绑定在一起的属性相同的端口可以像一个端口那样使用，通过聚合交换机之间的多条链路，可以在不进行硬件升级的前提下，增加交换机之间的带宽，并且使交换机之间的流量可以均衡分布到多条链路上，同时，多条链路还可以提供容错功能，在若干链路失效的情况下保证交换机之间的连通性。

4.1 链路聚合基础

4.1.1 链路聚合含义

假定图 4.1 中交换机 S1 和交换机 S2 只有 100Mb/s 端口，由于交换机之间不允许存在环路，如果需要提高交换机之间的带宽，不能简单地通过增加两台交换机之间的链路数量来提高交换机之间的带宽。链路聚合(Link Aggregation)(也称端口聚合)技术可以将多个端口聚合后作为单个端口使用。如图 4.1 所示，交换机 S1 和交换机 S2 之间三条 100Mb/s 链路通过链路聚合技术聚合在一起后，完全等同于一条 300Mb/s 链路，为了做到这一点，要求：

- 从聚合在一起的多个端口中的某个端口接收到的广播帧，不会从聚合在一起的其他端口中转发出去；
- 其中一台交换机可以将传输给另一台交换机的流量均衡地分布到聚合在一起的多个端口连接的多条链路上；
- 聚合在一起的多个端口中，若干端口，或端口连接的链路发生故障，只会影响交换机之间的带宽，不会影响两台交换机之间的连通性。

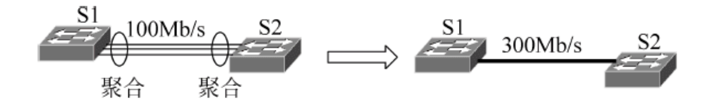


图 4.1　链路聚合含义

每一台交换机中聚合在一起作为单个端口使用的一组端口称为聚合组，由于每一台交换机允许同时存在多个不同的聚合组，需要用标识符标识不同的聚合组，这种用于标识聚合组的标识符称为聚合组标识符。互连两台交换机聚合组的一组链路称为链路聚合组，同样需要用链路聚合组标识符标识不同的链路聚合组。需要强调的是，聚合组只涉及单个交换

机，聚合链路组涉及一组链路互连的两个交换机，当然，一组链路两端设备除了交换机，还可以是路由器和终端，因此，将一组链路两端的设备统称为系统。

4.1.2 链路聚合方式

链路聚合方式有静态聚合和动态聚合两种，静态聚合方式需要手工在两台交换机上各自创建聚合组，并手工将交换机端口分配给聚合组，两台交换机上分配给各自聚合组的端口必须具有相同属性（如相同传输速率、相同通信方式等），互连交换机的链路两端必须是属于各自聚合组的端口，并且链路两端的端口必须都是开通端口。某台交换机分配给某个聚合组的端口不会监测链路另一端端口的属性和状态，因此，一旦发生某条链路两端端口的属性和状态不一致的情况，可能丢失经过该链路传输的 MAC 帧。

动态聚合方式通过链路聚合控制协议（Link Aggregation Control Protocol，LACP）动态分配聚合组中的端口，通过交换 LACP 报文相互监测链路另一端端口的状态和属性，当 LACP 监测到链路两端端口具有相同属性和状态时，链路两端端口才被加入到各自聚合组，当 LACP 监测到链路两端端口的属性和状态不一致时，链路两端端口将从各自聚合组中删除。

4.1.3 端口属性

属于同一聚合组的端口下述属性需要保持一致。

- STP 配置。端口路径距离、STP 报文格式、端口连接的链路类型（点对点链路或共享链路）、是否边缘端口、端口是否关闭等。
- VLAN 配置。端口属于的 VLAN 必须相同。
- 端口配置。端口传输速率、端口通信方式（半双工或全双工）、端口类型（接入端口或共享端口）、端口连接的链路类型（双绞线或光纤）。

4.2 链路聚合机制

4.2.1 功能组成

链路聚合功能组成如图 4.2 所示，总体上分为聚合器和链路聚合控制两大块，聚合器实现将一组聚合在一起的端口当作一个端口使用的功能，链路聚合控制实现将一组相同属性和状态的端口聚合在一起的功能。

1. 聚合控制

聚合控制模块用于确定聚合在一起的一组端口，存在两种用于确定聚合在一起的一组端口的机制，一是手工配置，完全由管理员确定属于每一个聚合组的端口，端口和聚合组之间的绑定关系是静态的，属于特定聚合组的每一个端口不监测链路另一端端口的属性和状态，有可能发生链路两端聚合组不匹配的问题，即属于同一个聚合组的一组端口所连的一组链路的另一端端口可能分布在不同的聚合组中。二是通过链路聚合控制协议动态分配属于每一个聚合组的端口，但这种动态分配并非不需要手工配置，通常通过手工配置确定属于特

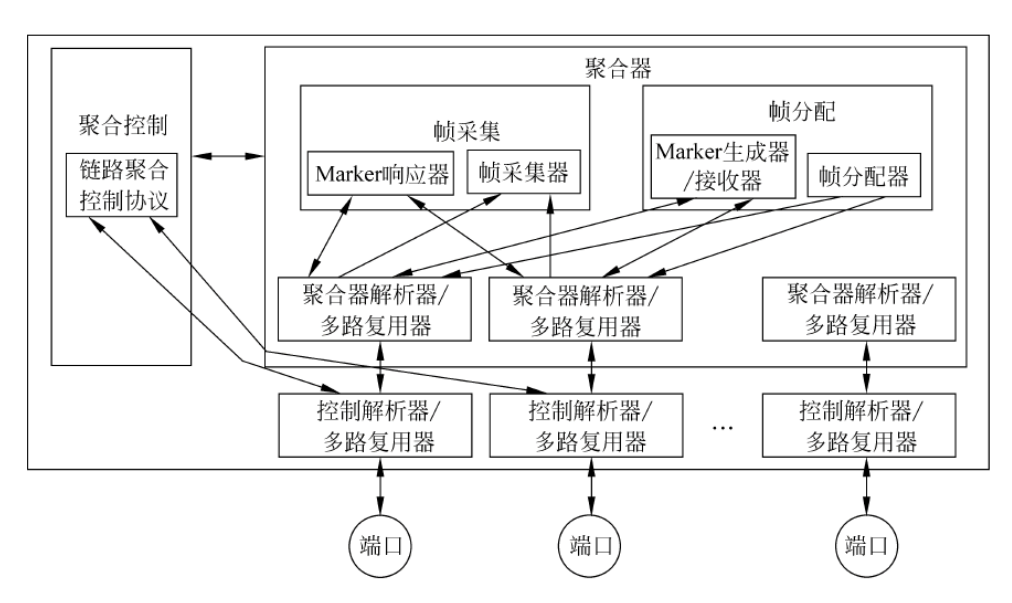


图 4.2 链路聚合功能组成

定聚合组的端口，然后由链路聚合控制协议监测链路另一端端口的属性和状态，属于某个特定聚合组的一组端口中只有由链路聚合控制协议保证其连接的一组链路的另一端端口属于同一个聚合组的这部分端口才能被激活，一旦链路一端端口的属性发生变化，链路聚合控制协议自动将链路两端端口从激活状态转换成关闭状态。

2. 控制解析器/多路复用器

控制解析器/多路复用器的功能一是将来自聚合器和链路控制模块的 MAC 帧复合在一起，通过交换机端口输出；二是从交换机端口接收到 MAC 帧时，区分出处理 MAC 帧的实体，将 MAC 帧分别送往聚合器或链路控制模块。这里的链路控制模块通常是链路聚合控制协议实体。控制解析器/多路复用器通过 MAC 帧的目的地址及净荷中的子类型字段值确定处理该 MAC 帧的实体。如果 MAC 帧的目的地址是组地址 01-80-C2-00-00-02，净荷中的子类型字段值为 1，该 MAC 帧被送往 LACP 实体。

3. 聚合器

聚合器主要由帧采集模块、帧分配模块和聚合器解析器/多路复用器组成。

1）帧分配模块

每个聚合组包含一组端口，对于交换机而言，属于同一聚合组的一组端口等同于单个端口，因此，转发表中，将聚合组作为单个输出端口，交换机 MAC 帧转发进程只能根据 MAC 帧的目的地址确定聚合组，由帧分配模块确定输出 MAC 帧的交换机端口。帧分配模块将通过聚合组输出的 MAC 帧根据流量均衡原则，分配到属于同一个聚合组的所有端口，并且保证这种分配过程不会引发经过以太网传输的 MAC 帧的错序。由于 MAC 帧经过端口输出时，可能需要在端口的输出队列中等待一段时间，等待时间长短与经过该端口的流量有关，因此，先到达交换机、从端口 X 输出的 MAC 帧，可能比后到达交换机、从端口 Y 输出的 MAC 帧后离开交换机。由于以太网端到端传输路径是唯一的，对于某台交换机，相同两端

的 MAC 帧的输入/输出端口是不变的，因此，经过以太网传输的 MAC 帧是不会错序的。但如果对于某台交换机，相同两端的 MAC 帧的输入/输出端口是变化的，就有可能导致相同两端的 MAC 帧经过以太网传输后错序，因此，帧分配算法必须保证将相同两端的 MAC 帧分配到聚合组中的同一个端口。分配算法可以基于以下 MAC 帧首部中的字段值和 MAC 帧封装的 IP 分组首部中的字段值分配端口。

- 源 MAC 地址。根据 MAC 帧的源 MAC 地址分配端口，有着相同源 MAC 地址的 MAC 帧被分配到聚合组中的同一个端口。
- 目的 MAC 地址。根据 MAC 帧的目的 MAC 地址分配端口，有着相同目的 MAC 地址的 MAC 帧被分配到聚合组中的同一个端口。
- 源 MAC 地址和目的 MAC 地址。根据 MAC 帧的源 MAC 地址和目的 MAC 地址分配端口，有着相同源 MAC 地址和目的 MAC 地址的 MAC 帧被分配到聚合组中的同一个端口。
- 源 IP 地址。根据 MAC 帧封装的 IP 分组的源 IP 地址分配端口，所有封装了有着相同源 IP 地址的 IP 分组的 MAC 帧被分配到聚合组中的同一个端口。
- 目的 IP 地址。根据 MAC 帧封装的 IP 分组的目的 IP 地址分配端口，所有封装了有着相同目的 IP 地址的 IP 分组的 MAC 帧被分配到聚合组中的同一个端口。
- 源 IP 地址和目的 IP 地址。根据 MAC 帧封装的 IP 分组的源 IP 地址和目的 IP 地址分配端口，所有封装了有着相同源 IP 地址和目的 IP 地址的 IP 分组的 MAC 帧被分配到聚合组中的同一个端口。

选择端口分配机制必须充分考虑网络结构，否则可能无法通过端口聚合技术实现线性增加设备间带宽的目的。图 4.3 中，路由器 R1 的多个交换端口聚合后作为单个物理接口，为物理接口分配单一的 IP 地址，该 IP 地址成为所有交换机 S1 连接的终端的默认网关地址，这些终端通过 ARP 地址解析过程解析默认网关地址对应的 MAC 地址时，获得相同的 MAC 地址。因此，所有发送给路由器 R1 的 MAC 帧有着相同的目的 MAC 地址，如果交换机 S1 使用基于 MAC 帧目的 MAC 地址选择聚合组中端口的端口分配机制，导致这些终端发送给路由器 R1 的 MAC 帧选择了聚合组中的同一个端口。同样，由于路由器 R1 发送给这些终端的 MAC 帧有着相同的源 MAC 地址，因此，路由器 R1 也不能使用基于 MAC 帧源 MAC 地址选择聚合组中端口的端口分配机制，否则无法通过端口聚合技术实现线性增加交换机 S1 和路由器 R1 之间带宽的目的。由于所有经过交换机 S2 转发的 MAC 帧有着相同的源 MAC 地址和目的 MAC 地址，因此，交换机 S2 不能使用基于 MAC 帧中 MAC 地址选择聚合组中端口的端口分配机制。

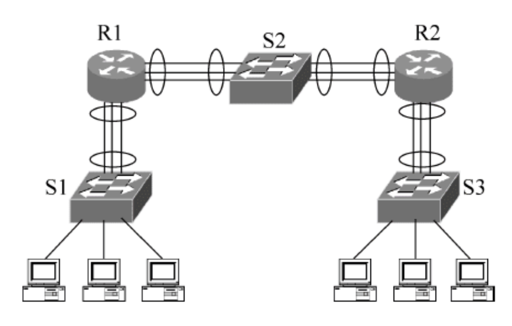


图 4.3 采用链路聚合技术的网络结构

2）帧采集模块

帧采集模块将通过属于同一聚合组中端口接收到的 MAC 帧提交给转发进程，由于端口分配机制保证相同两端的 MAC 帧经过交换机中相同的输入/输出端口，因此，帧采集模块无须对通过不同端口接收到的 MAC 帧排序。

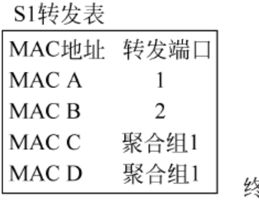
3）聚合器解析器/多路复用器

为了保证相同两端的 MAC 帧经过聚合链路传输后不会错序，要求相同两端的 MAC 帧经过聚合链路中的同一物理链路传输。聚合链路的容错性要求在某条物理链路无法继续提供传输服务时，将原来分配给该物理链路的流量分散到其他物理链路上，但必须保证经过其他物理链路传输的 MAC 帧不能先于已经分配给该物理链路传输的 MAC 帧到达聚合链路的另一端。存在两种实现这一功能的机制，一是设置定时器，在发生物理链路切换后，经过规定时间才开始通过新的物理链路传输 MAC 帧，规定时间保证分配给旧物理链路的 MAC 帧或者已经到达链路的另一端，或者已经丢失。二是采用 Marker 协议，在发生物理链路切换后，帧分配模块中的 Marker 协议实体生成一个 Marker 请求帧，并通过旧的物理链路传输 Marker 请求帧，对端的帧采集模块中的 Marker 协议实体接收到 Marker 请求帧后，立即发送一个 Marker 响应帧，当帧分配模块中的 Marker 协议实体接收到 Marker 响应帧，表明分配给旧物理链路的 MAC 帧已经完成传输，分配模块可以通过新的物理链路传输 MAC 帧。这样，经过控制解析器/多路复用器分流出的 MAC 帧存在两种类型，一种是数据帧，一种是 Marker 请求或响应帧，聚合器解析器/多路复用器的功能就是从经过控制解析器/多路复用器分流出的 MAC 帧中解析出 Marker 请求或响应帧，并将 Marker 请求或响应帧发送给 Marker 协议实体。聚合器解析器/多路复用器通过 MAC 帧的目的地址及净荷中的子类型字段值确定处理该 MAC 帧的实体。如果 MAC 帧的目的地址是组地址 01-80-C2-00-00-02，净荷中的子类型字段值为 2，该 MAC 帧被送往 Marker 协议实体。

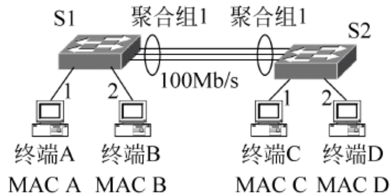
4.2.2　交换机通过聚合组转发 MAC 帧过程

图 4.4 所示网络结构能够正常工作的前提是：①已经在交换机 S1 和交换机 S2 中各自创建一个聚合组；②交换机 S1 和交换机 S2 已经完成对聚合组的端口分配；③链路聚合组两端端口各自属于同一个聚合组。这种情况下，聚合组对于交换机 S1 和交换机 S2 等同于单个端口。当交换机通过属于聚合组的某个端口接收到 MAC 帧后，在转发表中创建一项转发项，该转发项将该 MAC 帧的源 MAC 地址与聚合组绑定在一起。当交换机通过检索转发表发现某个 MAC 帧的输出端口是聚合组时，将该 MAC 帧提交给和该聚合组绑定的聚合器，聚合器中的帧分配器根据配置的端口分配机制在属于聚合组的端口中确定用于输出该 MAC 帧的端口，并把该 MAC 帧提交给输出端口。

S1转发表



MAC地址	转发端口
MAC A	1
MAC B	2
MAC C	聚合组1
MAC D	聚合组1



S2转发表

MAC地址	转发端口
MAC A	聚合组1
MAC B	聚合组1
MAC C	1
MAC D	2

图 4.4　网络结构

交换机广播某个 MAC 帧时，聚合组作为单个端口，广播帧同样由帧分配器根据配置的端口分配机制在属于聚合组的端口中确定用于输出该广播帧的端口，因此，广播帧也只从属于聚合组的单个端口输出。从属于聚合组的某个端口接收到的广播帧，不能从属于同一聚

合组的其他端口输出,只能从不属于该聚合组的所有其他端口广播出去。

无论是地址学习、转发操作,还是广播,聚合组对于交换机等同于单个端口。需要通过聚合组输出的 MAC 帧(包括目的地址是广播地址和组地址的 MAC 帧)由聚合器的帧分配器选择单个属于聚合组的端口输出。通过属于聚合组的某个端口接收到的 MAC 帧(包括目的地址是广播地址和组地址的 MAC 帧)只能从不属于该聚合组的其他端口输出,不能从属于该聚合组的其他端口输出。

4.2.3 链路聚合组生成过程

交换机通过聚合组接收和发送 MAC 帧前,必须做到:①创建聚合组;②将端口分配给聚合组;③测试分配给同一聚合组的一组端口所连接的链路的另一端端口是否属于同一个聚合组;④将聚合组中激活的端口与聚合器绑定。

1. 创建聚合组

交换机默认状态下不存在聚合组,因此,需要手工创建聚合组,每一个交换机允许创建的聚合组数量是有限的。创建聚合组时,需要为新创建的聚合组分配标识符,一般情况下,用数字标识不同的聚合组。

2. 分配端口

属于同一聚合组的端口必须具有相同属性,如传输速率、通信方式、所连接的传输媒体类型、端口所属的 VLAN 等,目前大多数交换机只允许连接点对点全双工链路的端口聚合在一起,因此,只有采用全双工通信方式的端口才允许分配给某个聚合组。一台交换机中,具有相同属性的端口很多,并不能将所有具有相同属性的端口自动分配给某个聚合组,聚合组和端口之间的绑定关系需要通过手工配置指定,因此,针对已经创建的每一个聚合组,需要通过手工配置指定分配给该聚合组的端口。

3. 激活端口

分配给某个聚合组中的端口存在两种类型,一是选中(Selected)端口,二是没有选中(Unselected)端口,只有选中端口才与聚合器绑定,才真正允许通过该端口发送、接收 MAC 帧。没有选中端口不和聚合器绑定,端口所连链路不能传输 MAC 帧。如果采用静态链路聚合方式,则分配给某个聚合组的端口的类型通过手工配置确定,端口类型是固定的,与该端口所连链路另一端端口的属性无关。如果采用动态链路聚合方式,则分配给聚合组的一组端口中只有满足下述条件的端口才是选中端口。

① 这些端口所连链路的另一端端口属于同一个聚合组。

② 如果为聚合组设置了最大选中端口数量,满足条件①的端口数量小于聚合组最大选中端口数量。

为了能够从分配给聚合组的一组端口中确定选中端口,并将其绑定到与该聚合组关联的聚合器,LACP 完成下述功能。

1) 分配系统标识符和端口标识符

图 4.4 中的交换机 S1 和交换机 S2 都是系统,系统标识符由分配给系统的优先级和系

统的 MAC 地址组成，系统标识符较小的系统具有较高的优先级。每一个端口的端口标识符由分配给端口的优先级和端口号组成，端口标识符较小的端口具有较高的优先级。

2）操作键

能够绑定到与某个聚合组关联的聚合器的一组端口，具有以下特点。

- 手工分配给该聚合组；
- 端口属性相同；
- 端口处于转发状态。

这样一组端口分配一个相同的整数，这个整数称为操作键，能够绑定到某个聚合器的一组端口必须具有相同的操作键，同样，这一组端口所连接链路的另一端端口也必须分配相同的操作键。

3）确定选中端口

聚合链路互连的两个系统中，具有较高优先级的系统为主系统，另一个系统为从系统，分配给某个聚合组的所有端口可以交换 LACP 报文，报文中给出发送端口的端口标识符、发送端口所在系统的系统标识符和分配给发送端口的操作键。主系统从具有下述特性的一组端口中选择一个优先级最高的端口作为主端口。

- 该端口已经分配与 X 聚合组关联的特定操作键；
- 该端口所连链路的另一端端口已经分配与 Y 聚合组关联的特定操作键。

将主端口绑定到与 X 聚合组关联的聚合器中，同时通过向对端端口发送 LACP 报文，要求对端系统将主端口所连链路的另一端端口绑定到与 Y 聚合组关联的聚合器中。将两端端口所在系统的系统标识符和分配给两端端口的操作键作为聚合链路的标识符。

以后，只有当某条链路两端端口所在系统的系统标识符和分配给端口的操作键与聚合链路的标识符相同时，该链路的两端端口才能绑定到聚合链路两端的聚合器上。如果为某个聚合组设置了最大选中端口数量 N，则主系统根据优先级选择优先级较高的 N 个满足选中端口条件的端口作为选中端口。

4.3　链路聚合控制协议

4.3.1　LACP 简介

1. 基本功能

LACP 的功能是动态构建链路聚合组，链路聚合组构建过程中，两端系统中创建聚合组和将端口分配到聚合组的过程通常通过手工配置完成，LACP 需要完成的功能是确定一组两端端口属于相同聚合组的链路，并把这一组链路两端的端口绑定到端口所属聚合组所关联的聚合器上，并实时监测这一组链路两端端口属性的变化过程，和两端聚合组中其他端口属性的变化过程，在某个链路聚合组中动态增加或删除某条链路。

为了做到这一点，LACP 需要定期交换链路两端端口属性和状态，如果监测到链路两端端口属性与某个已经建立的链路聚合组匹配，就将该链路增加到该链路聚合组。如果监测到某条已经加入某个链路聚合组的链路的两端端口属性发生变化，使得该链路两端端口属

性不再与该链路聚合组匹配，就将该链路从该链路聚合组中删除。

2. 端口模式

LACP 将端口模式分为主动(Active)和被动(Passive)两种模式，主动模式端口定期发送 LACP 报文，被动模式端口只有接收到对端发送的 LACP 报文，才回送 LACP 报文。因此，链路两端端口至少有一个端口的模式是主动模式。

4.3.2 LACP 报文格式

LACP 报文格式如图 4.5 所示，LACP 将发送报文的端口称为 Actor，将接收报文的端口称为 Partner，Actor 所在系统称为 Actor 系统，Partner 所在系统称为 Partner 系统，系统标识符、端口标识符和端口操作键的含义已经在 4.2.3 节中做了介绍。Actor 和 Partner 状态各占一个字节，每一位的含义如下。

目的 MAC 地址		类型	
0180C2000002	源 MAC 地址	8809	净荷
			子类型
			Actor 系统标识符
			Actor 操作键
			Actor 端口标识符
			Actor 状态
			Partner 系统标识符
			Partner 操作键
			Partner 端口标识符
			Partner 状态

图 4.5　LACP 报文格式

bit0 为 1，表示端口处于 Active 模式，bit0 为 0 表示端口处于 Passive 模式。

bit1 为 1，表示端口处于长溢出方式，每 30 秒发送一个 LACP 报文，溢出时间为 90 秒，bit1 为 0 表示端口处于短溢出方式，每秒发送一个 LACP 报文，溢出时间为 3 秒。

bit2 为 1，表示端口可以和其他端口聚合，bit2 为 0 表示端口不能和其他端口聚合。

bit3 为 1，表示端口已经绑定到对应聚合器，bit3 为 0 表示端口没有绑定到对应聚合器。

bit4 为 1，表示已经使能端口的帧采集器，bit4 为 0 表示没有使能端口的帧采集器。

bit5 为 1，表示已经使能端口的帧分配器，bit5 为 0 表示没有使能端口的帧分配器。

bit6 为 1，表示端口的信息来自 LACP，bit6 为 0 表示端口信息来自手工配置。

bit7 为 1，表示端口接收到的 LACP 报文已经超时，bit7 为 0 表示端口接收到的 LACP 报文没有超时。

4.3.3 LACP 工作过程

1. 将链路接入到链路聚合组的过程

LACP 开始工作前，必须完成系统的初始配置，对于图 4.6 所示的系统 A 和系统 B，初

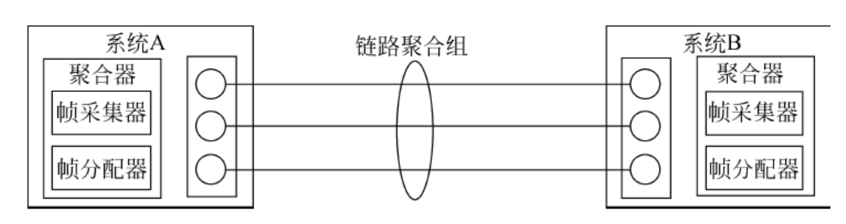


图 4.6　链路聚合过程

始配置如下。

- 分别在系统 A 和系统 B 中创建聚合组 X 和聚合组 Y，并且分别将三个端口分配给聚合组 X 和聚合组 Y，系统 A 和系统 B 中分别分配给聚合组 X 和聚合组 Y 的三个端口有着相同操作键 B 和操作键 C。操作键 B 和操作键 C 根据端口所属的聚合组和端口属性生成；
- 分别为系统 A 和系统 B 分配优先级，系统 A 和系统 B 根据分配的优先级（或默认优先级）和自己的 MAC 地址生成系统标识符 IDA 和 IDB，假定 IDA<IDB；
- 属于聚合组 X 和聚合组 Y 的三个端口，分别根据分配的端口优先级（或默认优先级）和端口编号生成端口标识符 X1～X3 和 Y1～Y3，假定 X1<X2<X3；
- 将 6 个端口的模式配置为主动模式（或端口 X1、端口 X2、端口 X3 为主动模式，端口 Y1、端口 Y2 和端口 Y3 为被动模式）。

完成上述配置后，链路两端端口开始交换 LACP 报文，图 4.7 是端口 X1 和端口 Y1 交换 LACP 报文过程。首先由端口 X1 发送 LACP 报文<IDA，B>，表明端口 X1 所在系统的系统标识符是 IDA，端口 X1 的操作键是 B。端口 Y1 接收到端口 X1 发送的 LACP 报文后，向端口 X1 发送 LACP 报文<IDB，C，IDA，B>，LACP 报文除了端口 X1 的信息外，还给出端口 Y1 所在系统的系统标识符 IDB 和端口 Y1 的操作键 C。端口 X1 接收到端口 Y1 发送的 LACP 报文后，得知端口 X1 和端口 Y1 是链路两端端口。由于系统 A 是主系统，且端口 X1 的端口标识符最小，将端口 X1 绑定到和聚合组 X 关联的聚合器，使能聚合器中的帧采集器。端口 X1 在发送给端口 Y1 的 LACP 报文中把 Actor 状态中 bit3 和 bit4 设置为 1。端口 Y1 接收到该 LACP 后，将端口 Y1 绑定到与聚合组 Y 关联的聚合器，使能聚合器中的帧采集器和帧分配器，创建链路聚合组<IDA，B，IDB，C>。<IDA，B，IDB，C>是用两端端口所在系统的系统标识符和两端端口的操作键构成的链路聚合组标识符。端口 Y1 向端口 X1 发送 LACP 报文，并在 LACP 报文中把 Actor 状态中 bit3、bit4 和 bit5 设置为 1，端口 X1 接收到该 LACP 报文后，使能聚合器中的帧分配器，创建链路聚合组<IDA，B，IDB，C>。此时，两端系统之间已经可以通过链路聚合组传输 MAC 帧。

当端口 X2 和端口 Y2 之间、端口 X3 和端口 Y3 之间相互交换 LACP 报文后，发现两端端口所在系统的系统标识符和两端端口的操作键与已经建立的链路聚合组<IDA，B，IDB，C>匹配，将互连端口 X2 与端口 Y2、端口 X3 与端口 Y3 的链路加入该链路聚合组。将某条链路加入该链路聚合组意味着已经完成将链路两端端口各自绑定到聚合组 X 和聚合组 Y 关联的聚合器，并向链路另一端端口发送将 Actor 状态中 bit3、bit4 和 bit5 设置为 1 的 LACP 报文的过程。

只要链路正常，链路两端端口之间定期交换 LACP 报文，端口定时器不会溢出，链路和链路聚合组之间的绑定关系一直维持。



<IDA,B ,(b3=b4=b5=1),IDB,C,(b3=b4=b5=1)>表示如下：
Actor标识符：IDA
Actor操作键：B
Actor状态：bit3、bit4和bit5设置为1
Partner标识符：IDB
Partner操作键：C
Partner状态：bit3、bit4和bit5设置为1

图 4.7 链路两端端口之间 LACP 报文交换过程

2. 链路从链路聚合组分离的过程

两种情况导致链路从链路聚合组分离，一是长时间接收不到链路另一端端口发送的 LACP 报文，二是链路两端端口中至少有一个端口的操作键发生改变。

如果某个端口在长溢出时间段(默认时间为 90s)内一直没有接收到链路另一端端口发送的 LACP 报文，将导致定时器溢出，该端口将定时器溢出时间设置为短溢出时间段(默认时间为 3s)，并在发送给链路另一端端口的 LACP 报文中将 Actor 状态中的 bit1 设置成 0，要求链路另一端端口立即发送 LACP 报文。如果在短溢出时间段内仍然一直没有接收到链路另一端端口发送的 LACP 报文，将再次导致定时器溢出，表明链路或链路另一端端口发生问题，该端口将和聚合器分离。

如果端口的操作键发生改变，该端口将和绑定的聚合器分离，并在发送给链路另一端端口的 LACP 报文中给出新的端口操作键，并将 Actor 状态中的 bit2 设置成 0。链路另一端系统同样将该端口和绑定的聚合器分离。链路两端通过交换 LACP 报文开始将该链路加入到新的链路聚合组的过程。改变端口操作键的原因主要有以下这些：

- 通过手工配置改变了端口所属的聚合组；
- 端口属性发生改变，如通信方式由全双工变为半双工；
- 端口所属的 VLAN 发生改变。

习题

4.1 可以聚合在一起的链路必须具有什么共性？

4.2 链路聚合组生成过程由哪些阶段组成？

4.3 LACP 的作用是什么?

4.4 静态配置链路聚合组会导致什么问题?

4.5 聚合组和链路聚合组有什么区别和联系?

4.6 网络结构如图 4.3 所示,交换机 S2 需要配置几个聚合组?每一个聚合组需要配置何种端口分配机制?解释原因。

第5章 路由器和网络互连

目前的现状是多种类型网络共存，并独立发展，每一种网络已经解决连接在该网络上的两个终端之间的通信功能，如以太网已经实现连接在以太网上的两个终端之间的 MAC 帧传输过程，问题是，任何一种网络都无法独立实现全球任何两个终端之间的通信功能，只有把多种有着不同的适用范围、不同的功能特性的网络互连在一起，才能实现全球任何两个终端之间的通信功能，这就需要一种新的独立于任何类型网络的端到端分组传输协议——IP，一种新的用于互连不同类型网络的设备——路由器。

5.1 网络互连

5.1.1 网络互连需要解决的问题

实现不同种类传输网络（异构网络）的互连就是实现两个连接在不同种类的传输网络上的终端之间的端到端通信。图 5.1 是实现以太网和公共交换电话网（Public Switched Telephone Network，PSTN）互连的互连网络结构，PSTN 是由 PSTN 交换机和互连 PSTN 交换机的物理链路构成的网络，通过呼叫连接建立过程建立两个结点之间的点对点语音信道。图 5.1 所示互连网络的核心是路由器，路由器为了实现以太网和 PSTN 的互连，必须具有以太网端口和 PSTN 端口，能够通过以太网端口实现和连接在以太网上的终端之间 MAC 帧的传输，同样，能够在 PSTN 端口和连接在 PSTN 上的终端之间建立语音信道，并通过语音信道实现数据传输。为了通过以太网实现路由器以太网端口和其他连接在以太网上的终端之间的 MAC 帧传输，以太网端口需要分配 MAC 地址。同样，为了通过呼叫连接建立过程在路由器 PSTN 端口和其他连接在 PSTN 上的终端之间建立语音信道，需要为 PSTN 端口分配电话号码。这样，必须经过路由器中继，才能实现连接在不同传输网络上的两个终端之间的通信过程，因此，图 5.1 中终端 A 至终端 B 的传输路径由两部分组成，一是终端 A 至路由器以太网端口的以太网交换路径，该交换路径根据以太网交换机中的转发表和路由器以太网端口的 MAC 地址（MAC R）确定。二是路由器 PSTN 端口和终端 B 之间的语音信道，该语音信道通过路由器 PSTN 端口和终端 B 之间的呼叫连接建立过程建立。实际的数据传输过程应该这样，终端 A 将需要传输给终端 B 的数据封装在以 MAC A 为源 MAC 地址、以 MAC R 为目的 MAC 地址的 MAC 帧中，通过以太网将该 MAC 帧传输给路由器以太网端口。路由器从 MAC 帧中分离出数据，确定数据的目的终端（终端 B）位于

PSTN 端口连接的 PSTN,通过呼叫连接建立过程建立路由器 PSTN 端口和终端 B 之间的点对点语音信道,然后通过点对点语音信道将数据传输给终端 B。虽然点对点语音信道不需要寻址,但为了检测出数据传输过程中发生的错误,仍然需要将数据封装成帧,帧中除了数据字段,还须有检错码字段。因此,路由器首先将数据封装成适合点对点语音信道传输的帧格式,然后通过点对点语音信道将帧传输给终端 B。但图 5.1 所示的互连网络结构实现这样的传输过程还存在一些问题。

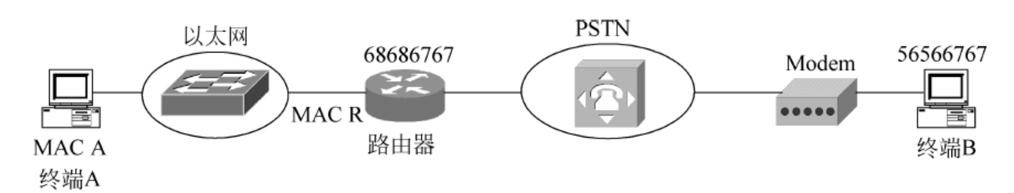


图 5.1 互连网络结构

1. 源终端路径选择问题

终端 A 向终端 B 传输数据前,需要获取终端 B 的地址,对于 PSTN,标识终端的地址是电话号码,如图 5.1 中终端 B 的电话号码 56566767。终端 A 即使获得了终端 B 的电话号码,如何确定终端 A 至终端 B 的传输路径?对于图 5.1 所示的互连网络结构,终端 A 如何根据终端 B 的电话号码确定实现以太网和 PSTN 互连的路由器及路由器以太网端口的 MAC 地址?

2. 目的终端标识问题

终端 A 传输给终端 B 的数据封装在 MAC 帧中,路由器通过以太网端口接收到 MAC 帧后,需要通过呼叫连接建立过程建立路由器 PSTN 端口和终端 B 之间的语音信道,但 PSTN 的呼叫连接建立过程需要被叫和主叫的电话号码,路由器如何根据接收到的 MAC 帧及封装在 MAC 帧中的数据确定数据的目的终端及目的终端的电话号码?

3. 数据封装问题

适合以太网传输的数据封装形式是 MAC 帧,适合点对点语音信道传输的数据封装形式是点对点协议(Point-to-Point Protocol,PPP)帧,这是两种截然不同的封装形式,路由器根本无法根据 MAC 帧或 PPP 帧中包含的端到端传输的数据实现这两种封装形式的相互转换。

5.1.2 信件投递过程的启示

图 5.2 是将信件从南京投递到长沙的示意图,首先,寄信人用来传递信息的信纸被封装为信件,信封上写上寄信人和收信人地址,然后将信件提交给南京邮局,南京邮局根据收信人地址:长沙,确定信件的下一站:上海,由于南京至上海采用公路运输系统,因此,信件被封装为适合公路运输系统的形式:信袋,而且信袋上用车次 3536 表明运输该信袋的车辆及始站与终站。由于上海至长沙采用航空运输系统,上海首先从信袋中提取出信件,然后将其

封装成适合航空运输系统的形式：信盒，信盒上用航班号 AU765 表明运输该信盒的飞机及始站与终站。信件经过南京至上海的公路运输系统和上海至长沙的航空运输系统这两阶段的运输服务到达目的地：长沙。通过如图 5.2 所示的信件投递过程，可以得出以下启示：

（1）不同运输系统有着不同的封装信件的形式和标识始站与终站的方式；

（2）信件上收信人和寄信人地址是统一的，和实际提供运输服务的运输系统标识始站与终站的方式无关；

（3）信件是一种标准的封装形式，和实际提供运输服务的运输系统封装信件的形式无关；

（4）南京根据信件上的收信人地址确定下一站：上海，同样，上海也是根据信件上的收信人地址确定下一站：长沙，信件在南京至长沙的传输过程中是不变的；

（5）由实际的运输系统提供当前站至下一站的运输服务。

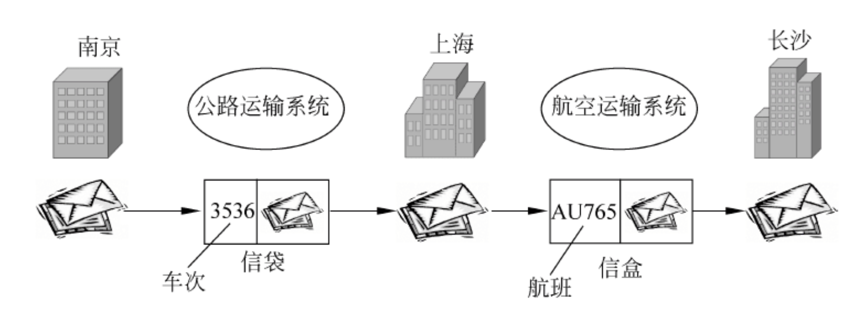


图 5.2　信件传输过程

5.1.3　端到端传输思路

根据图 5.2 所示的信件投递过程，可以引申出实现数据端到端传输的思路：

（1）定义一种和具体传输网络无关的、统一的终端地址格式：IP 地址。

（2）定义一种和具体传输网络无关的、统一的数据封装格式：IP 分组，并在 IP 分组中用 IP 地址标识数据的源和目的终端。

（3）如图 5.3 所示，端到端传输路径由终端、各种不同类型的传输网络和互连不同类型传输网络的路由器组成，终端和路由器称为跳，数据从源终端开始传输，首先传输给源终端至目的终端传输路径上的第一跳路由器，该路由器和源终端连接在同一个传输网络上，对应源终端为下一跳。数据端到端传输过程由多个传输阶段组成，每一个传输阶段实现数据从当前跳至和当前跳连接在同一传输网络的下一跳的传输过程，这种传输方式称为逐跳传输，每一跳通过 IP 分组中目的终端的 IP 地址确定下一跳。

（4）通过实际的传输网络实现 IP 分组当前跳至下一跳的传输过程，IP 分组经过实际传输网络传输时，须封装成实际传输网络对应的格式。

图 5.3 给出的数据端到端传输过程所涉及的步骤及功能如下：

（1）必须为终端 A、终端 B 及路由器分配统一的 IP 地址：IP A、IP B 和 IP R；

（2）终端 A 和路由器必须通过路由表给出用于根据终端 B 的 IP 地址（IP B）确定下一跳的信息；

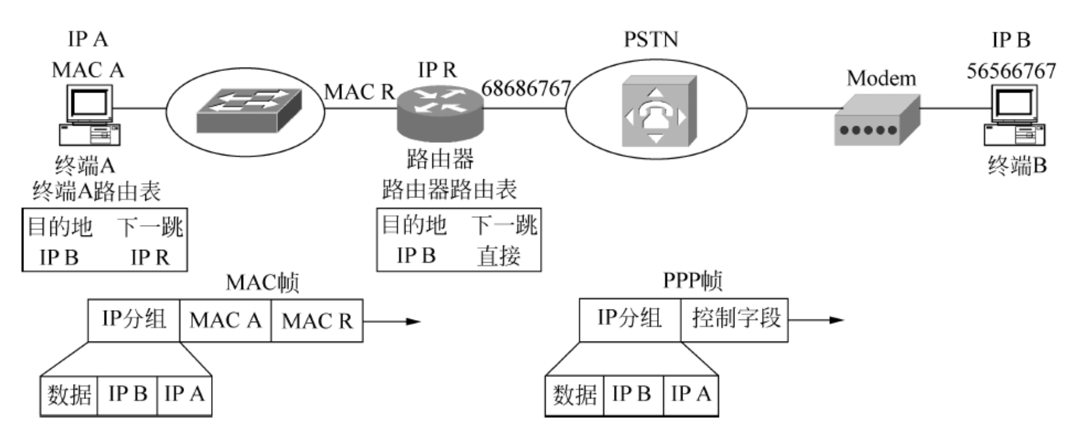


图 5.3 端到端数据传输过程

(3) 终端 A 将传输给终端 B 的数据封装成 IP 分组格式,并在 IP 分组中给出终端 A 和终端 B 的 IP 地址,端到端传输过程中,IP 分组是不变的;

(4) 终端 A 根据 IP 分组中终端 B 的 IP 地址确定下一跳:路由器,并获得路由器的 IP 地址:IP R,终端 A 必须根据路由器的 IP 地址确定连接终端 A 和路由器的传输网络——以太网,获得路由器以太网端口的 MAC 地址,将 IP 分组封装成 MAC 帧,经过以太网将 MAC 帧传输给路由器;

(5) 路由器从 MAC 帧中分离出 IP 分组,根据 IP 分组中终端 B 的 IP 地址确定下一跳:终端 B(路由器路由表中用"直接"表明下一跳就是目的终端自身),根据终端 B 的 IP 地址确定连接终端 B 和路由器的传输网络——PSTN,并获得终端 B 的电话号码,通过呼叫连接建立过程建立路由器和终端 B 之间的点对点语音信道,将 IP 分组封装成适合点对点语音信道传输的格式:PPP 帧,并经过点对点语音信道将 PPP 帧传输给终端 B;

(6) 终端 B 从 PPP 帧中分离出 IP 分组,再从 IP 分组中分离出终端 A 传输给它的数据。

5.1.4 IP 实现网络互连机制

1. 路由器和 IP 实现端到端传输的过程

从图 5.3 所示的端到端传输过程,可以得出网际协议(Internet Protocol,IP)实现网络互连的机制:

(1) 规定了统一的且与传输网络地址标识方式无关的 IP 地址格式,所有接入互连网络的终端必须分配一个唯一的 IP 地址,同时,由于每一个终端都和实际传输网络相连,还需具有实际传输网络相关的地址,如以太网的 MAC 地址,为了区分,将 IP 地址称为逻辑地址,将实际传输网络相关的地址称为物理地址。

(2) 规定了统一的且与传输网络数据封装格式无关的 IP 分组格式,端到端传输的数据必须封装成 IP 分组,每一跳通过 IP 分组携带的目的终端 IP 地址确定下一跳。

(3) 对应每一个目的终端,每一跳必须建立用于确定通往该目的终端的传输路径上的下一跳的信息,该信息被称为路由项,它主要由两部分组成:目的终端 IP 地址和通往该目的终端的传输路径上的下一跳的 IP 地址。对应多个不同目的终端的路由项集合,称为路由表。

(4) 必须由单个传输网络连接当前跳和下一跳,能够根据下一跳 IP 地址确定连接当前跳和下一跳的传输网络,解析出下一跳的传输网络相关地址,即物理地址,能够将 IP 分组封装成传输网络要求的帧格式,并通过互连当前跳和下一跳的传输网络实现 IP 分组当前跳至下一跳的传输过程。

(5)IP 分组经过逐跳传输,实现源终端至目的终端的传输过程。

2. 路由表和路由项

实现互连网络端到端传输过程的关键有两点,一是源终端能够根据目的终端 IP 地址确定通往目的终端传输路径上的第一跳路由器,源终端至目的终端传输路径经过的路由器能够根据目的终端 IP 地址确定通往目的终端传输路径上的下一跳路由器。二是互连当前跳和下一跳的网络能够实现 IP 分组当前跳至下一跳的传输过程。如果互连当前跳和下一跳的网络是以太网,实现第二关键点需要完成以下三个步骤:①根据下一跳的 IP 地址解析出下一跳连接以太网接口的 MAC 地址;②将 IP 分组封装成以当前跳连接以太网接口的 MAC 地址为源 MAC 地址、下一跳连接以太网接口的 MAC 地址为目的地址的 MAC 帧;③经过以太网实现该 MAC 帧当前跳连接以太网接口至下一跳连接以太网接口的传输过程。前面有关以太网的章节已经详细讨论了完成第②步骤、第③步骤的方法和过程。实现第一关键点的核心是路由表,对应每一个目的终端地址,需要路由表给出通往该目的终端传输路径上的下一跳结点的 IP 地址。因此,路由表中的路由项格式是＜目的终端 IP 地址,下一跳结点 IP 地址＞,如果当前跳和目的终端连接在同一个网络,则当前跳至目的终端传输路径不存在其他路由器,下一跳结点 IP 地址用“直接”表示。如果当前跳通往一组目的终端的传输路径有着相同的下一跳结点,只需为这一组目的终端设置一项路由项＜表示这一组目的终端的 IP 地址,下一跳结点 IP 地址＞。

IP 分组端到端传输过程中,源终端将 IP 分组传输给第一跳路由器的过程,或者当前路由器根据目的 IP 地址和路由表确定下一跳路由器,并将 IP 分组传输给下一跳路由器的过程称为间接交付。源和目的终端位于同一个传输网络,或者路由器根据目的 IP 地址和路由表确定的下一跳是目的终端本身(目的 IP 地址匹配的路由项中的下一跳为“直接”),源终端或路由器通过直接连接的传输网络将 IP 分组传输给目的终端的过程称为直接交付。

5.1.5 数据报 IP 分组交换网络

为了简化互连网络端到端数据传输过程,在建立互连网络端到端传输路径时,可以将互连终端和路由器及路由器间互连的传输网络虚化成链路,将图 5.4(a)所示的互连网络结构简化为图 5.4(b)所示的由路由器互连多条链路而成的数据报分组交换网络,在图 5.4(b)所示的数据报分组交换网络中,路由器就是 IP 分组交换机,路由表就是转发表,每一个 IP 分组都是独立的数据报,路由器根据路由表和 IP 分组携带的目的终端 IP 地址实现 IP 分组的转发操作。由于可以将互连网络作为数据报 IP 分组交换网络进行分析、处理,因而常常用数据报 IP 分组交换网络(简称 IP 网络)来称呼互连网络。在 IP 网络中,传输网络的功能等同于逻辑链路,用于实现终端和 IP 分组交换机(路由器),及两个相邻 IP 分组交换机之间的 IP 分组传输,因而将传输网络的分组格式称为帧,将传输网络中的分组交换设备,如以太网交换机,称为链路层设备(第二层设备)。但 OSI 体系结构中定义的用于互连分组交换机的

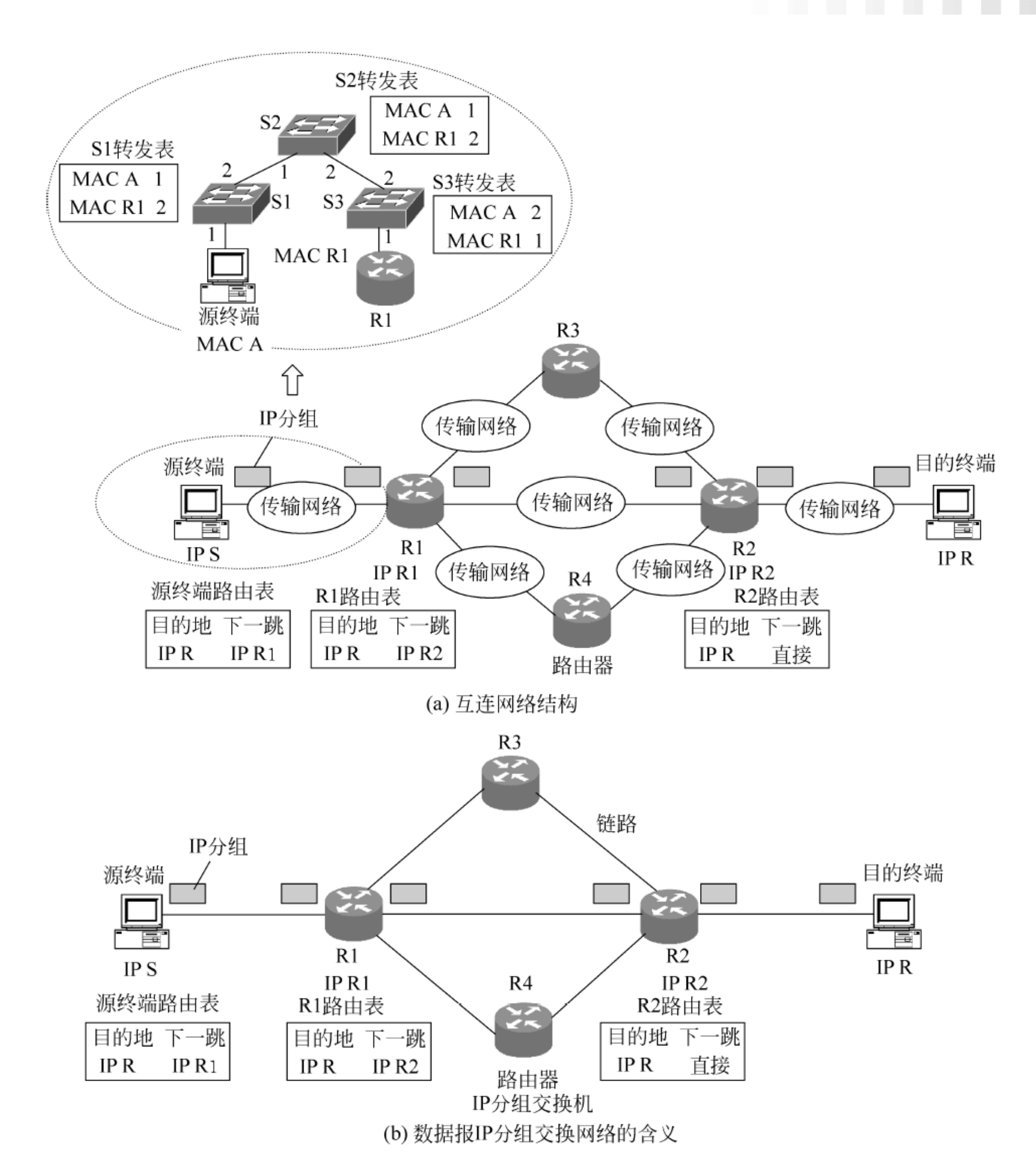


图 5.4　数据报 IP 分组交换网络

链路是点对点物理链路或广播物理链路(多点接入物理链路),因此,OSI 体系结构定义的网络应该是由终端、物理链路、分组交换机组成的,与图 5.4(a)所示的由终端、传输网络和用于互连传输网络的路由器组成的互连网络是不同的。对于图 5.4(a)所示的互连网络,端到端传输路径实际上由两层传输路径组成,一是 IP 层传输路径,由源终端、中间路由器和目的终端组成,如图 5.4(a)中源终端至目的终端传输路径:源终端→路由器 R1→路由器 R2→目的终端。二是传输网络中当前跳至下一跳的传输路径,如果连接源终端和路由器 R1 的传输网络是如图 5.4(a)中展开的交换式以太网,则源终端至路由器 R1 传输路径就是由源终端、中间以太网交换机和路由器 R1 组成的交换路径:源终端→以太网交换机 S1→以太网交换机 S2→以太网交换机 S3→路由器 R1。因此,和互连网络有关的内容由三部分组成,

它们分别是：IP、路由协议和 IP over X 技术。IP 详细规定了 IP 地址格式和 IP 分组格式。路由协议通过为每一个路由器建立路由表实现建立 IP 层传输路径的功能。IP over X(X 指特定的传输网络)技术实现根据下一跳 IP 地址解析出下一跳连接 X 传输网络的端口的物理地址，和 IP 分组经 X 传输网络从当前跳传输给下一跳的功能。

5.1.6 路由器结构

路由器从本质上讲是 IP 分组转发设备，根据 IP 分组首部中的目的终端 IP 地址完成 IP 分组从输入端口至输出端口的转发过程。图 5.5 是路由器功能结构，从功能上可以把路由器分成三部分：路由模块、线卡和交换模块。

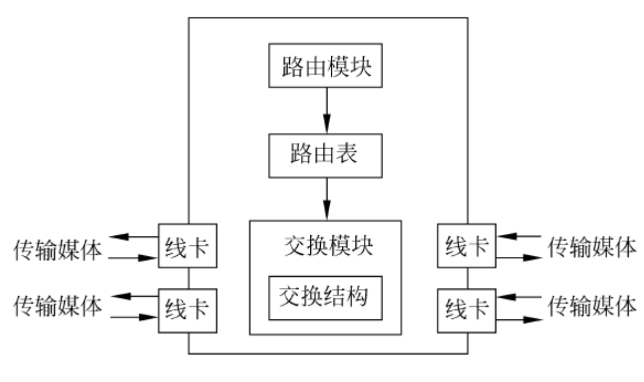


图 5.5 路由器结构

1. 路由模块

路由模块负责运行路由协议，生成路由表，在路由表中给出到达互连网络中任何一个终端的传输路径，当然，由于 IP 分组是逐跳转发，路由器的路由表中只需给出通往互连网络中某个终端的传输路径上的下一跳路由器的地址。由于生成路由表的过程比较复杂，因此，路由模块的核心部件通常是 CPU，大部分功能由软件实现。除了生成路由表，路由模块也承担一些其他的管理功能。

2. 线卡

线卡负责连接外部传输媒体，并通过传输媒体连接传输网络，如连接以太网的线卡通过双绞线或光纤连接以太网交换机。线卡通过端口连接传输媒体，不同类型的传输媒体对应不同类型的端口，如连接 5 类双绞线的端口(俗称电端口)和连接光纤的端口(俗称光端口)。线卡除了实现和传输网络的物理连接，还需要按照所连接的传输网络的要求完成 IP 分组的封装和分离操作。封装操作将 IP 分组封装成适合通过传输网络传输的链路层帧格式，如以太网的 MAC 帧。分离操作和封装操作相反，从链路层帧中分离出 IP 分组。线卡进行接收操作时，从所连接的传输媒体接收到的物理层信号(如曼彻斯特编码流)中分离出链路层帧(如 MAC 帧)，并从链路层帧中分离出 IP 分组。发送操作时，将 IP 分组封装成链路层帧(如 MAC 帧)，将链路层帧转换成物理层信号(如曼彻斯特编码流)后，发送到传输媒体。线卡的每一个端口还需配置输入、输出队列，用于存储无法及时处理的 IP 分组。

3. 交换模块

当线卡从某个端口接收到的物理层信号中分离出 IP 分组，就将该 IP 分组发送给交换模块。交换模块用 IP 分组的目的终端 IP 地址检索路由表，找到输出端口，并把 IP 分组发送给输出端口所在的线卡。随着端口的传输速率越来越高，如 10Gb/s 的以太网端口，端口每秒接收、发送的 IP 分组数量越来越大，对于 10Gb/s 的以太网端口，在极端情况下(假定 IP 分组的长度为 46B，MAC 帧的长度为 64B)，端口每秒接收、发送的 IP 分组数量 $=10\times10^9/(64\times8)=19.53$M IP 分组/s(19.53Mpps)，当路由器多个端口都线速接收、发送 IP 分组时，交换模块的处理压力将变得很大，因此，通常用称为交换结构的专用硬件来完成 IP 分组从输入端口到输出端口的转发处理。由于存在从多个输入端口输入的 IP 分组需要从同一个输出端口输出的情况，即使交换结构能够支持所有端口线速接收、发送 IP 分组，输出端口也需要设置输出队列，用输出队列来临时存储那些无法及时输出的 IP 分组。

路由器是实现不同类型的传输网络互连的关键设备，它一方面通过路由模块建立到达任何终端的传输路径，另一方面，在确定下一跳结点的 IP 地址后，完成下一跳结点 IP 地址到下一跳结点所连接的传输网络所对应的链路层地址的转换，并将 IP 分组封装成传输网络要求的链路层帧格式，通过传输网络传输给下一跳结点。

5.2 网际协议

网际协议(Internet Protocol，IP)是实现连接在不同类型传输网络上的终端之间通信功能的基础，用于定义独立于传输网络的 IP 地址和 IP 分组格式。

5.2.1 IP 地址分类

在深入讨论 IP 地址前，需要说明一下，IP 地址不是终端或路由器的标识符，而是终端或路由器接口的标识符，就像地址不是房子的标识符，而只是门牌号一样。一栋房子如果有多个门，则有多个不同的门牌号，也就有多个不同的地址，但以这些地址为收信人地址的信件都能投递给该房子的主人。同样，终端或路由器允许有多个接口，每一个接口都有独立的标识符——IP 地址，但以这些 IP 地址为目的地址的 IP 分组都能到达该终端或路由器。接口是终端或路由器连接网络的地方，多数情况下，终端或路由器的每一个端口都连接独立的网络，这种情况下，接口就是端口。但在一些特殊情况下，一个端口可能同时连接多个不同的网络，或是多个端口连接同一网络，因而，一个端口可能对应多个不同的接口，或是多个端口对应同一个接口。下面章节将针对具体应用，对这些特殊情况进行讨论。由于每一个 IP 地址指向唯一的终端或路由器，因此，从这一点上讲，IP 地址确实有终端或路由器标识符的作用。

1. IP 地址分类方法

图 5.6 给出了 IP 地址的分类方法。一般情况所指的 IP 地址是指 IPv4 所定义的 IP 地址，它由 32 位二进制数组成，为了表示方便，将 32 位二进制数分成 4 个 8 位二进制数，每个 8 位二进制数单独用十进制表示(0～255)，4 个用十进制表示的 8 位二进制数用点分隔，如 32 位二

进制数表示的 IP 地址：01011101 10100101 11011011 11001001，表示成 93.165.219.201。

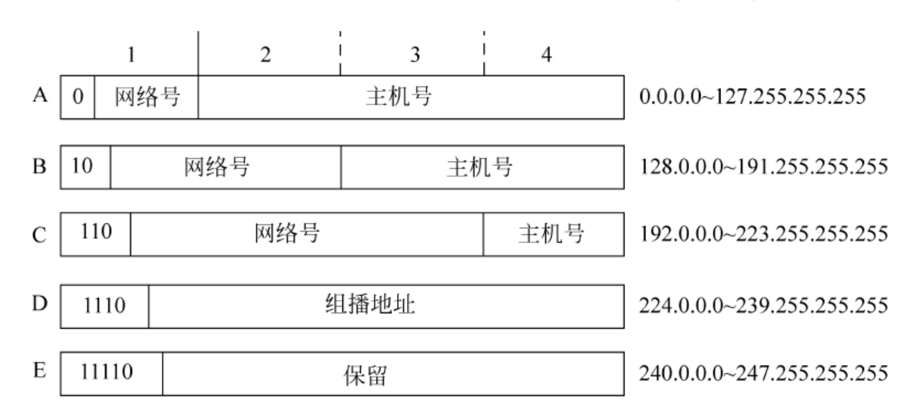


图 5.6 IP 地址分类方法

IP 是实现网络互连的协议，因此，用来标识互联网中终端设备的每一个 IP 地址由两部份组成：网络号和主机号。最高位为 0，表示是 A 类地址，用 7 位二进制数标识网络号，24 位二进制数标识主机号，A 类地址中网络号全 0 和全 1 的 IP 地址有特别用途，不能作为普通地址使用。0.0.0.0 表示 IP 地址无法确定，终端没有分配 IP 地址前，可以用 0.0.0.0 作为 IP 分组的源地址。127.X.X.X 是环回地址。所有类型的 IP 地址中，主机号全 0 和全 1 的 IP 地址也有特别用途，也不能作为普通地址使用。如网络号为 5 的 A 类 IP 地址的范围为 5.0.0.0～5.255.255.255，但 IP 地址 5.0.0.0 用于表示网络号为 5 的网络地址，而 IP 地址 5.255.255.255 作为在网络号为 5 的网络内广播的广播地址。A 类地址的范围是 0.0.0.0～127.255.255.255，但实际能用的网络号是 1～126，每一个网络号下允许使用的主机号＝$2^{24}-2$，由此可以看出，A 类地址适用于大型网络。

最高 2 位为 10，表示 B 类地址，用 14 位二进制数标识网络号，用 16 位二进制数标识主机号，能够标识的网络号为 2^{14}，每一个网络号下允许使用的主机号＝$2^{16}-2$。B 类地址的范围是 128.0.0.0～191.255.255.255，适用于大、中型网络。

最高 3 位为 110，表示是 C 类地址，用 21 位二进制数表示网络号，8 位二进制数表示主机号，能够标识的网络号为 2^{21}，每一个网络号下能够标识的主机号＝$2^{8}-2$。很显然，C 类地址只适用于小型网络。实际应用中并不使用 B 类和 C 类地址中网络号全 0 的 IP 地址。

A、B、C 三类地址都称为单播地址，用于唯一标识 IP 网络中的某个终端，但任何网络内都有一个主机号全 1 的地址作为该网络内的广播地址，这种广播地址不能用于标识网络内的终端，只能在传输 IP 分组时作为目的地址，表明接收方是网络内的所有终端。

每一个单播 IP 地址具有唯一的网络号，因此，对应唯一的网络地址，根据单播 IP 地址求出对应的网络地址的过程如下：根据该 IP 地址的最高字节值确定该 IP 地址的类型，根据类型确定主机号字段位数，清零主机号字段得到的结果就是该 IP 地址对应的网络地址。如 IP 地址 193.1.2.7 对应的网络地址为 193.1.2.0。

最高 4 位为 1110，表示是组播地址，用 28 位二进制数标识组播组，同一个组播组内的终端可以任意分布在 Internet 中，因此，组播组是不受网络范围影响的。有些组播地址有特殊用途，称为著名组播地址，下面就是一些常用的著名组播地址，这些组播地址表明接收端是同一网络内的特定结点。

224.0.0.1 表示网络中所有支持组播的终端和路由器。
224.0.0.2 表示网络中所有支持组播的路由器。
224.0.0.4 DVMRP 路由器。
224.0.0.5 表示网络中所有运行 OSPF 进程的路由器。
224.0.0.9 表示网络中所有运行 RIP 进程的路由器。
最高 5 位为 11110,表示是 E 类地址,目前没有定义。

32 位全 1 的 IP 地址称为受限广播地址,以受限广播地址为目的地址的 IP 分组在当前网络内广播,所有路由器均不转发以受限广播地址为目的地址的 IP 分组。特定网络内主机号全 1 的 IP 地址称为直接广播地址,也称为定向广播地址,以直接广播地址为目的地址的 IP 分组在由网络号指定的特定网络内广播,所有没有和该特定网络直接相连的路由器正常转发该 IP 分组。

2. 互连网络 IP 地址配置原则

互连网络配置 IP 地址的原则如下:

(1) 连接在同一传输网络上的终端必须配置具有相同网络号的 IP 地址,如连接在以太网上的终端 A 和终端 C;

(2) 每一个传输网络都有一个网络地址,如图 5.7 中以太网配置的网络地址 192.1.1.0 和 PSTN 配置的网络地址 192.1.2.0;

(3) 路由器的每一个接口都需配置 IP 地址,该 IP 地址对应的网络地址必须和分配给该接口所连的传输网络的网络地址相同,如图 5.7 中连接以太网接口配置的 IP 地址 192.1.1.254,其网络地址为 192.1.1.0,和以太网配置的网络地址相同。

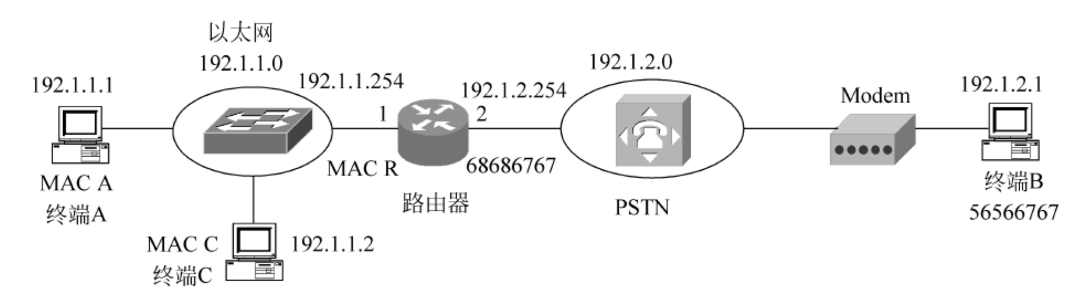


图 5.7 IP 地址配置

如果一个物理以太网被划分为多个 VLAN,则每一个 VLAN 就是一个独立的传输网络,不同 VLAN 须配置不同的网络地址,需要用路由器实现多个 VLAN 的互连。

5.2.2 IP 地址分层分类的原因和缺陷

1. 根据考号寻找考场的启示

图 5.8 假定考号由 6 位十进制数组成,其中高 3 位是考场号,低 3 位是座位号,同一考场的考号具有相同的考场号。每一个考场的考场号随机分配。用 751XXX 表示该考场的考场号是 751,座位号包括该考场内的全部座位号,因此,可用 751XXX 表示考场号是 751 的考场。

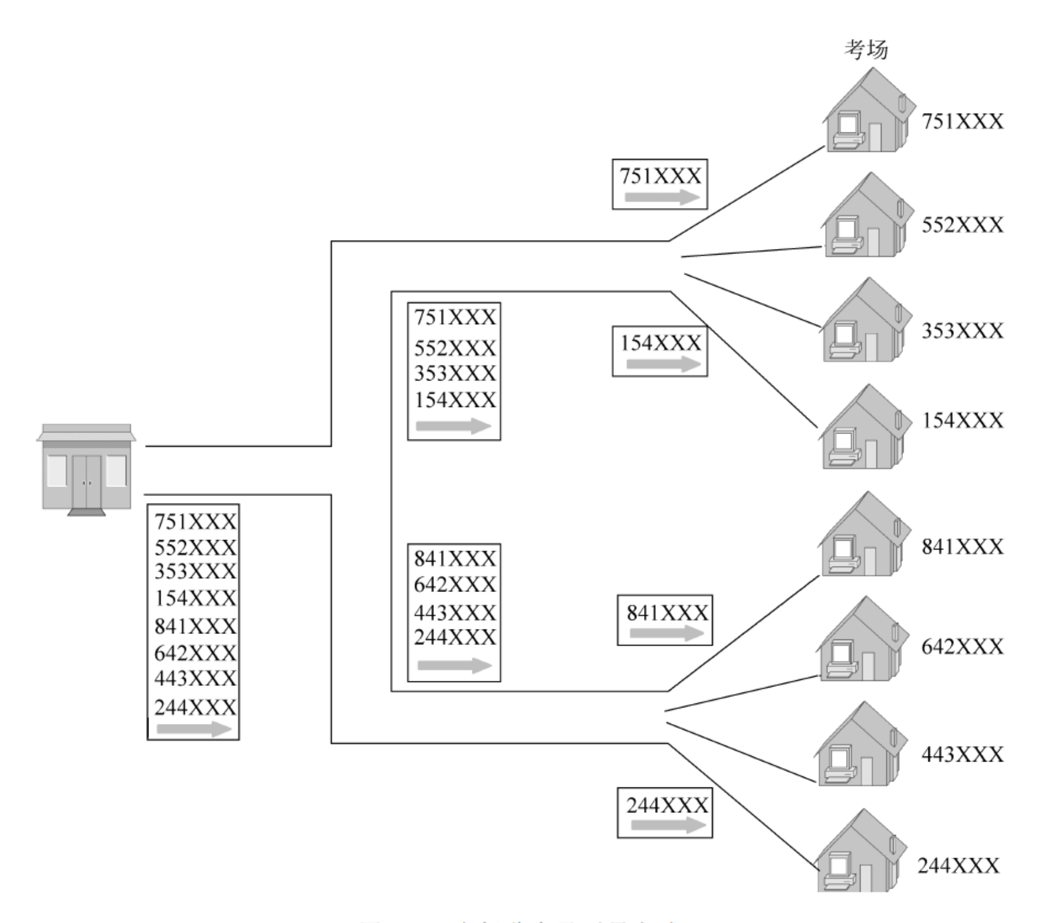


图 5.8 考场分布及引导方式

图中的考场指示方式必须保证能够将所有考号属于这 8 个考场的考生正确引导到考号指定的考场，每一个考生，在每一个路口，用考号的高 3 位比较路牌中各项的考场号，一旦考号中的考场号和路牌中某项的考场号相同，表示考号和该项匹配，考生将沿着该项给出的方向继续前进，直到正确到达考号指定考场，然后，再在考场内根据座位号找到正确的座位。由于考号分为两层，因此，路口路牌中的每一项只需给出特定考场的考场号及该考场的前进方向，无须为每一个考号设置前进方向。

两层结构减少了路牌中的项数，但当该单位设置多个考场时，路口路牌中的项数仍然偏多，有什么方式可以在两层结构不变的前提下，减少路牌中的项数？

如果考场号的分配方式如图 5.9 所示，符合以下分配原则：

(1) 最高位为 6 的考场号只分配给分布在该单位的考场，且该单位所有考场的考场号的最高位一定为 6；

(2) 分配给同一个区域中考场的考场号的次高位号必须是相同的，且分配给不同区域中考场的考场号的次高位必须不同。

根据上述分配原则分配考场号，可以进一步减少路口路牌的项数，但考生在不同路口用于匹配路牌中每一项的考号的位数是变化的，如单位入口用于匹配路牌中每一项的是考号的最高位，丁字路口用于匹配路牌中每一项的是考号的高 2 位，因此，必须在路口路牌中给出用于匹配路牌中每一项的考号位数。

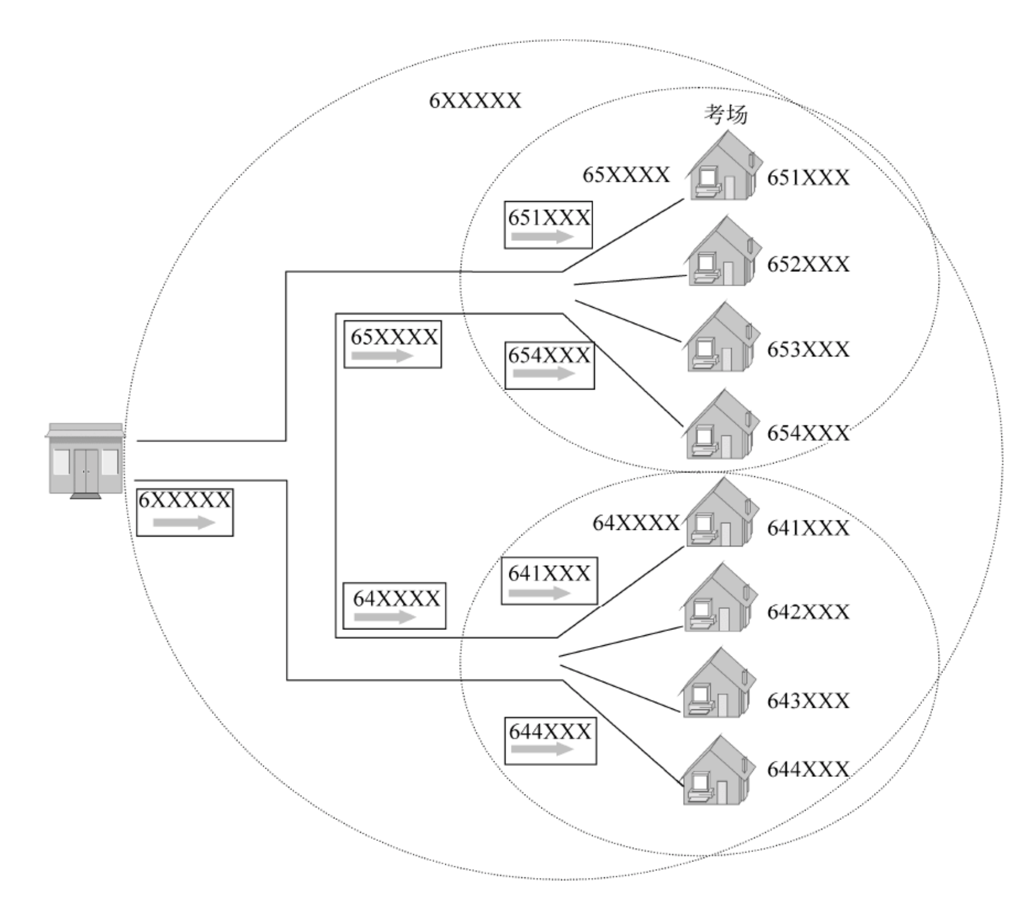


图 5.9　特定考场号分配规则下考场引导方式

根据两层考号寻找考号对应的考场座位的过程给出以下启示：

(1) 考生首先根据考号中考场号找到考场，然后，在考场寻找座位，因此，路口路牌只需指出通往每一个考场的路径；

(2) 通过限制考场号的分配，使得路口路牌可以用考场号的最高位或高 2 位指定分布在单位内的所有考场，或分布在单位内某个区域中的所有考场，以此减少路口路牌中的项数。

2. IP 地址分层分类的原因

IP 地址分层的目的也是希望用一项路由项指出通往该网络内所有终端的传输路径，如图 5.10 所示。需要说明的是：一般情况下，路由项格式是＜目的网络，下一跳路由器 IP 地址＞，目的网络字段给出目的终端所在网络的网络地址，这里为了讨论问题方便，假定图 5.10 中的路由器都直接用点对点物理链路相连，因此，路由项中只需给出转发端口就可确定通往目的终端的传输路径上的下一跳路由器，路由项格式变为＜目的网络，转发端口＞。但如果连接两个路由器的是类似以太网这样的传输网络，仅知道转发端口并不能确定通往目的终端的传输路径上的下一跳路由器，必须给出下一跳路由器的 IP 地址。

IP 地址为 192.1.1.1 的终端寻找通往 IP 地址为 192.2.1.1 的终端的传输路径的过程和图 5.8 中考生从单位入口开始根据考号寻找对应的考场座位的过程十分相似，每一个路

由器就像路口，而路由表就像路口路牌，转发端口就是前进方向，路由器根据目的终端的 IP 地址确定 IP 分组转发端口的过程如下：

(1) 求出目的终端 IP 地址对应的网络地址 N；

(2) 用 N 逐项比较路由表中每一项路由项的目的网络字段，如果和其中一项的目的网络字段值相同，表示目的终端的 IP 地址和该路由项匹配，通过该路由项指定的转发端口输出 IP 分组；

对于路由器 1，首先求出目的终端 IP 地址 192.2.1.1 对应的网络地址 192.2.1.0，然后逐项比较路由表中每一项路由项的目的网络字段值，结果和路由项＜192.2.1.0，端口 2＞的目的网络字段值相同，通过端口 2 输出该 IP 分组。两个终端之间传输路径上的所有路由器依次转发，最终将 IP 分组送达 IP 地址为 192.2.1.1 的终端。

IP 地址分类的原因是不同单位的网络规模是不同的，有些单位的网络规模很大，可以采用 B 类，甚至 A 类地址，有些单位的网络规模较小，可以采用 C 类地址，使得 IP 地址分配更贴近实际需要。

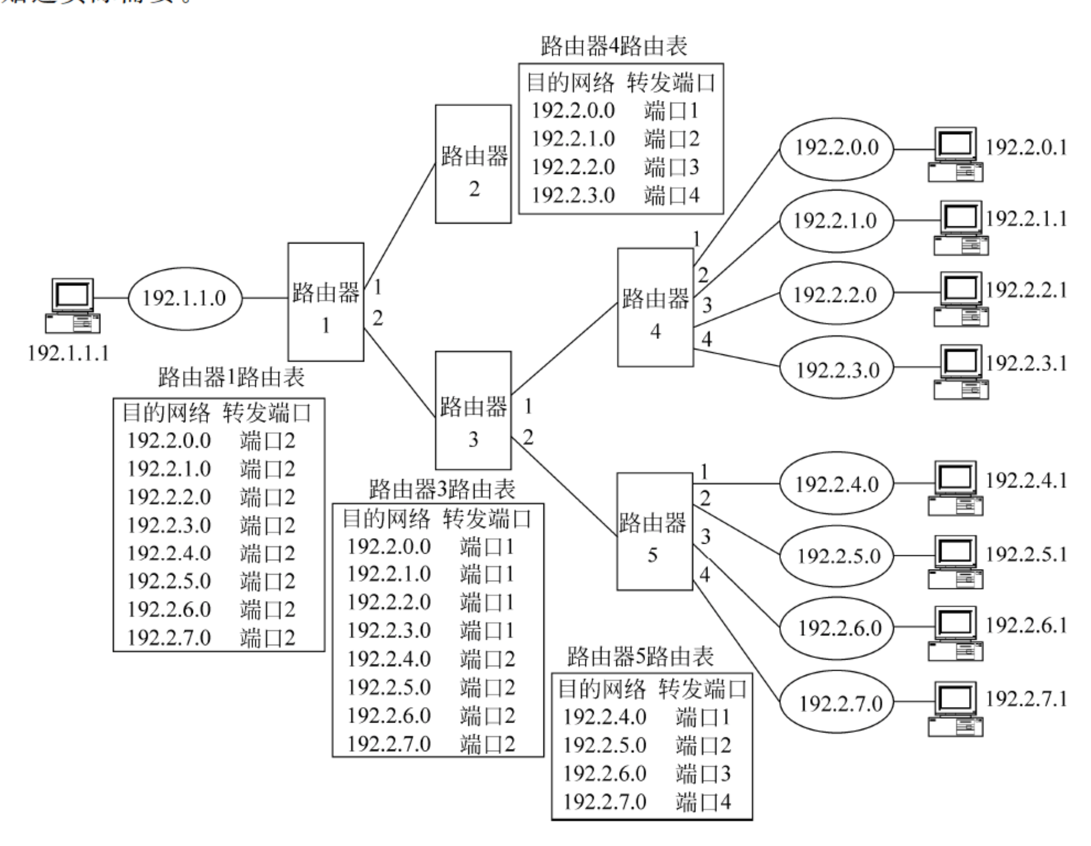


图 5.10 固定地址结构的路由方式

3. 固定地址结构的缺陷

图 5.10 所示的路由方式中，路由表必须为每一个网络设置一项路由项，这将使得路由表中的路由项数目非常庞大，由于大量网络之间的传输路径都需要经过核心路由器，因此核心路由器路由项数目庞大的问题尤为严重。路由项数目庞大，一是占用大量的存储空间，二是使得路由器根据目的终端 IP 地址确定下一跳路由器的转发处理变得耗时。

当初设计 IP 地址时,将 IP 单播地址分成 A、B、C 三类的目的是为了使 IP 地址能够适应不同规模的网络,地址分配能够更加贴近实际需要,避免浪费。但实际应用中,IP 地址浪费的问题依然十分严重,尽管将地址分成 A、B、C 三类,但许多网络规模介于两类地址之间,如具有 1000 个终端的网络,C 类地址不够,B 类地址又很浪费,因此,20 世纪 90 年代中期就出现 IP 地址短缺问题,而出现 IP 地址短缺问题的主要原因是大量 A 类和 B 类地址空间被浪费。

解决上述问题的关键是能够动态改变 IP 地址中网络号的位数,如图 5.9 所示的根据考号寻找考场座位的过程,如果不同路由器中和同一目的终端 IP 地址匹配的路由项的网络地址中的网络号的二进制位数是可变的,同样可以通过限制网络的网络号分配,用网络号的最高若干位表示网络号连续的一组网络号。更进一步,如果网络号的二进制位数不再固定,可以动态设置,可以根据实际网络规模申请主机号的位数,如 1000 台主机,可以申请一个 10 位主机号、22 位网络号的 IP 地址块,这样,传统的分类不再存在,网络号和主机号位数根据实际情况动态设置,这种思路就是无分类编址(Classless InterDomain Routing,CIDR,直译是无分类域间路由)。

5.2.3 无分类编址

1. 无分类编址机制

无分类编址方式下,32 位 IP 地址中标识网络号和主机号的二进制位数是可变的,这样做,消除了 IP 地址的分类,也解决了因为分类带来的种种问题。但必须提出一种用于指明 IP 地址中作为网络号的二进制位数的方法。无分类编址通过子网掩码(更确切的名称应该是网络掩码,但子网掩码已经成为习惯称呼)指明 IP 地址中作为网络号的二进制位数。子网掩码也是一个 32 位的二进制数,和 IP 地址的表示方法一样,也用 4 个点分隔的十进制数表示,每个十进制数表示 8 位二进制数,如 255.0.0.0,展开成二进制表示为 11111111 00000000 00000000 00000000。子网掩码中为 1 的二进制数对应 IP 地址中作为网络号的二进制数。5.1.1.2/255.0.0.0 表示 IP 地址是 5.1.1.2,对应的子网掩码是 255.0.0.0,如果将子网掩码展开成二进制表示,则只有高 8 位二进制数为 1,其余为 0,这就意味着 IP 地址的高 8 位为网络号,低 24 位为主机号。同样,5.1.1.2/255.255.255.0 表示 IP 地址的高 24 位为网络号,低 8 位为主机号。目前还有一种更直接的表示方式是直接给出 IP 地址中作为网络号的二进制位数,如 5.0.0.0/8、5.1.0.0/16、192.2.0.0/21 等。更简单的表示方式是省略 IP 地址中低位连续的 0,如 5.0.0.0/8 可以表示成 5/8,5.1.0.0/16 可以表示成 5.1/16。

图 5.9 中 6XXXXX 并不是表示一个 1 位考场号,5 位座位号的超大考场,而是考场号最高位为 6 的一组考场号的表示方式,同样,用 N 位网络前缀表示一组最高 N 位相同的连续网络号,网络前缀的表示方式和前面表示网络号的方式相同,可以用子网掩码或数字指定 32 位 IP 地址中网络前缀的位数,但网络前缀和网络号的含义不同,它只是用来表示具有相同网络前缀的一组 IP 地址,这一组 IP 地址称为 CIDR 地址块,可能由分配给不同网络的 IP 地址组成。<网络前缀,主机号>的 IP 地址结构完全取消了原先定义的 A、B、C 三类 IP 地址的概念,因而被称为无分类编址。N 位网络前缀的 CIDR 地址块可以分配给单个网络,这

种情况下,N 位网络前缀就是该网络的网络号。也可以分配给多个网络,这种情况下,N 位网络前缀只是用来确定 CIDR 地址块的 IP 地址范围。

2. 聚合路由项

图 5.10 中,路由器 1 对应 8 个网络的路由项有这样的特点:

① 8 个网络的网络号是连续的,所包含的 IP 地址范围为 192.2.0.0~192.2.7.255。这样范围内的 IP 地址的最高 21 位是相同的,而且地址范围包含了低 11 位(21+11=32)的全部 2^{11} 种组合。具有这样特性的 IP 地址块可以表示为网络前缀为 21 位的 CIDR 地址块,如果将<网络前缀,主机号>的 IP 地址结构中主机号字段为 0 的地址称为网络前缀地址的话,通过用最高 21 位为 1 的子网掩码和地址块中任何一个 IP 地址相与后获得的结果作为该 CIDR 地址块的网络前缀地址。以下是一个网络前缀地址的计算实例。

```
   11000000 00000010 00000001 00000001    192.2.1.1
&& 11111111 11111111 11111000 00000000    255.255.248.0
   -----------------------------------
   11000000 00000010 00000000 00000000    192.2.0.0
```

因此,网络前缀地址 192.2.0.0/21 包含的 IP 地址范围等于图 5.10 中 8 个网络所包含的 IP 地址范围。

② 8 个网络对应的路由项有着相同的转发端口,即有着相同的下一跳。

具有上述两个特点的 8 项路由项可以聚合为 1 项路由项<192.2.0.0/21,端口 2>。同样,路由器 3 中网络 192.2.0.0、192.2.1.0、192.2.2.0 和 192.2.3.0 与网络 192.2.4.0、192.2.5.0、192.2.6.0 和 192.2.7.0 所包含的 IP 地址块构成网络前缀为 22 位的 CIDR 地址块,这种地址块的网络前缀地址用最高 22 位为 1 的子网掩码和地址块中任何一个 IP 地址相与后获得,如计算网络 192.2.4.0、192.2.5.0、192.2.6.0 和 192.2.7.0 所包含的 CIDR 地址块对应的网络前缀地址的过程如下:

```
   11000000 00000010 00000101 00000111    192.2.5.7
&& 11111111 11111111 11111100 00000000    255.255.252.0
   -----------------------------------
   11000000 00000010 00000100 00000000    192.2.4.0
```

网络前缀地址 192.2.4.0/22 包含的 IP 地址范围等于网络 192.2.4.0、192.2.5.0、192.2.6.0 和 192.2.7.0 包含的 IP 地址范围。同样,网络前缀地址 192.2.0.0/22 包含的 IP 地址范围等于网络 192.2.0.0、192.2.1.0、192.2.2.0 和 192.2.3.0 包含的 IP 地址范围,且这两组网络对应的路由项有着相同的转发端口,因此,每一组网络对应的 4 项路由项可以聚合为 1 项路由项<192.2.0.0/22,端口 1>和<192.2.4.0/22,端口 2>。以此可以得出图 5.11 所示的采用无分类编址后的各路由器中的路由项。和图 5.10 比较,图 5.11 中的路由项数目大幅减少。路由项聚合后得出的由 N 位网络前缀指定的 CIDR 地址块是分配给一组网络号连续且高 N 位相同的网络的 IP 地址集合,只是通往这一组网络的传输路径有着相同的下一跳。

采用无分类编址后,路由器根据 IP 分组的目的终端 IP 地址确定下一跳路由器的过程如下:

(1) 根据路由项的网络前缀位数求出目的 IP 地址的网络前缀地址,然后和路由项的目的网络字段值比较,如果相同,表示该目的 IP 地址和该路由项匹配,通过路由项指定的转发

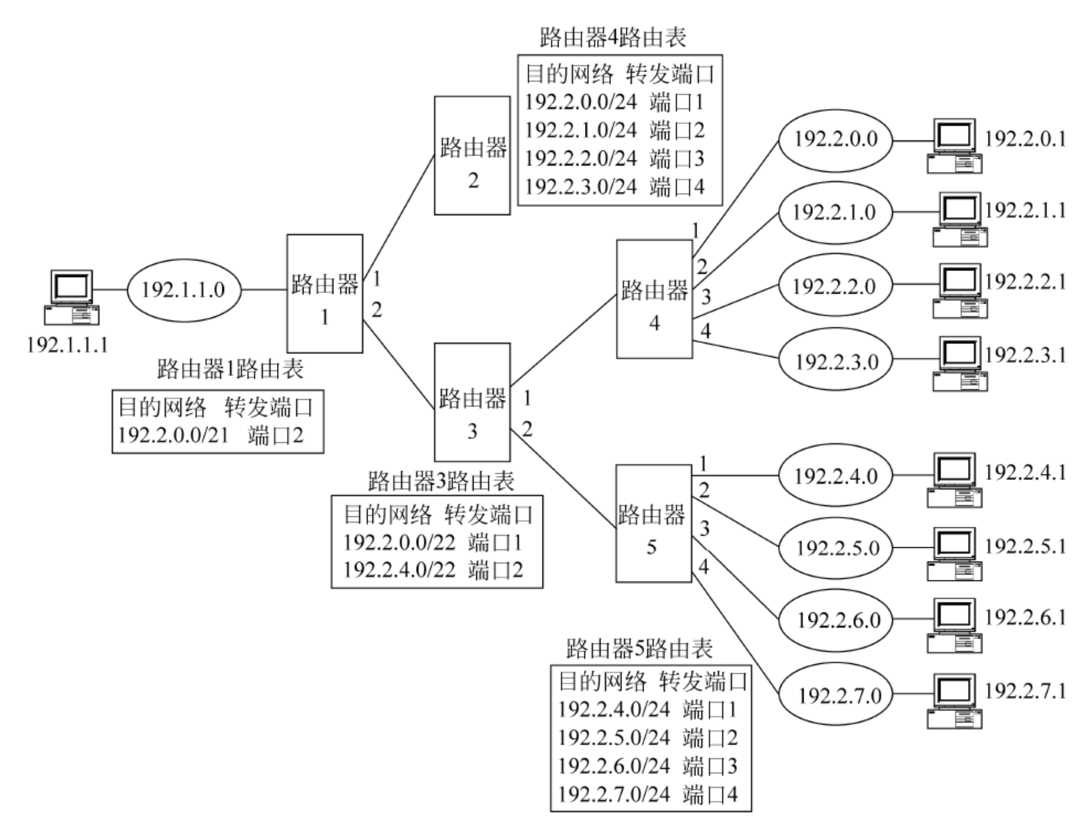


图 5.11 采用无分类编址后的路由项表示

端口输出该 IP 分组；

(2) 对路由表中的每一项路由项进行上述操作，直到找到匹配的路由项，或报错。

如果 IP 分组中的目的 IP 地址为 192.2.5.1，路由器 3 进行如下操作：

对于第一项路由项，由于网络前缀为 22 位，用最高 22 位为 1 的子网掩码(255.255.252.0)和 192.2.5.1 相与后获得对应的网络前缀地址。

```
    11000000 00000010 00000101 00000001    192.2.5.1
&&  11111111 11111111 11111100 00000000    255.255.252.0
   -------------------------------------
    11000000 00000010 00000100 00000000    192.2.4.0
```

由于目的 IP 地址 192.2.5.1 对应的网络前缀地址为 192.2.4.0/22，和路由项中的目的网络字段值 192.2.0.0/22 不同，因此，IP 分组的目的 IP 地址和第一项路由项不匹配。由于第二项路由项的网络前缀也是 22 位，目的 IP 地址对应的网络前缀地址同样为 192.2.4.0/22，它和第二项路由项中的目的网络字段值相同，因此，通过端口 2 输出该 IP 分组。

3. 任意划分子网

采用无分类编址的另一个好处是可以任意划分子网，假定某个单位有 120 台计算机，这些计算机被分成 6 组，其中第 1 组分配 20 台计算机，第 2 组分配 12 台计算机，第 3 组分配 45 台计算机，第 4 组分配 27 台计算机，第 5 组分配 5 台计算机，第 6 组分配 11 台计算机，这 6 组计算机属于 6 个子网，如何分配 IP 地址才能使得路由表中的路由项最少？为了使路由表中的路由项最少，分配给这些计算机的 IP 地址必须是连续的，因此，用最低 8 位二进制数

不同的 256 个 IP 地址作为这些计算机的 IP 地址，如 CIDR 地址块 192.1.2.0/24。最后 8 位二进制数的分配规则如下：

00000000～**00**111111(0～63)分配给第 3 组 45 台计算机，网络地址为 192.1.2.0/26；

01000000～**010**11111(64～95)分配给第 4 组 27 台计算机，网络地址为 192.1.2.64/27；

01100000～**011**11111(96～127)分配给第 1 组 20 台计算机，网络地址为 192.1.2.96/27；

10000000～**1000**1111(128～143)分配给第 2 组 12 台计算机，网络地址为 192.1.2.128/28；

10010000～**1001**1111(144～159)分配给第 6 组 11 台计算机，网络地址为 192.1.2.144/28；

10100000～**10100**111(160～167)分配给第 5 组 5 台计算机，网络地址为 192.1.2.160/29。

上述分配过程的思路如下，最多的是第 3 组有计算机 45 台，求出满足不等式 $2^N \geqslant 45+2$ 的最小 N，得出主机字段需要 6 位二进制数，8 位二进制数的表示范围可以分成 4 个 6 位二进制数的表示范围，高 2 位分别为 00、01、10 和 11。高 2 位为 00 的 64 个地址分配给第 3 组。第 4 组和第 1 组的计算机台数分别为 27 和 20，主机字段需要 5 位二进制数，高 2 位为 01 的 64 个地址可以分成高 3 位分别是 010 和 011 的 2 组 32 个地址，分别分配给第 4 组和第 1 组。第 2 组和第 6 组的计算机台数分别为 12 和 11，主机字段需要 4 位二进制数，高 2 位为 10 的 64 个地址可以分成高 4 位分别是 1000、1001、1010 和 1011 的 4 组 16 个地址，高 4 位分别为 1000 和 1001 的 2 组 16 个地址分别分配给第 2 组和第 6 组。高 4 位为 1010 的 16 个地址可以分成高 5 位分别是 10100 和 10101 的 2 组 8 个地址，将高 5 位为 10100 的 8 个地址分配给第 5 组。

图 5.12 中的路由器 R1 只需给出 1 项路由项＜192.1.2.0/24，192.1.1.1，1＞，表明只要目的 IP 地址高 24 位等于 192.1.2 的 IP 分组均转发给 IP 地址为 192.1.1.1 的下一跳路由器 R2，路由器 R2 对每一个子网均需给出 1 项路由项，目的网络字段值给出的 CIDR 地址块必须包含分配给该子网的全部 IP 地址。

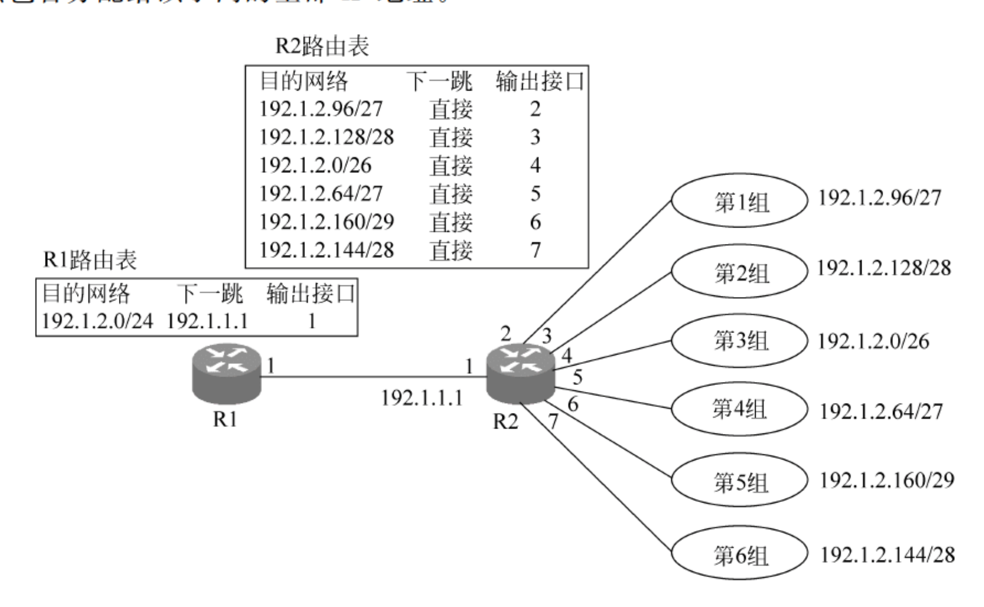


图 5.12 无分类编址任意划分子网过程

【例 5.1】 某网络的 IP 地址空间为 192.168.5.0/24，采用等长子网划分，子网掩码为 255.255.255.248，求子网数和每一个子网内可分配的 IP 地址数。

【解析】 根据子网掩码 255.255.255.248 得出网络号位数为 29 位，主机号位数为 3 位，192.168.5.0/24 IP 地址空间中，用 5 位(29－24＝5)作为子网号，求出子网数＝2^5＝32，子网内可分配的 IP 地址数＝2^3－2＝6。

【例 5.2】 某企业分配给人事部的 CIDR 地址块为 10.0.11.0/27，分配给企划部的 CIDR 地址块为 10.0.11.32/27，分配给市场部的 CIDR 地址块为 10.0.11.64/26，这三个 CIDR 地址块聚合后的 CIDR 地址块应是________。

A. 10.0.11.0/25　　　　B. 10.0.11.0/26

C. 10.0.11.64/25　　　　D. 10.0.11.64/26

【解析】 答案是 A，CIDR 地址块 10.0.11.0/27 表示的 IP 地址集合是：10.0.11.0(IP 地址低 8 位为 000 00000)～10.0.11.31(IP 地址低 8 位为 000 11111)，CIDR 地址块 10.0.11.32/27 表示的 IP 地址集合是：10.0.11.32(IP 地址低 8 位为 001 00000)～10.0.11.63(IP 地址低 8 位为 001 11111)，两个 CIDR 地址块聚合后的 IP 地址集合是：10.0.11.0(IP 地址低 8 位为 00 000000)～10.0.11.63(IP 地址低 8 位为 00 111111)，实际上就是 CIDR 地址块 10.0.11.0/26。CIDR 地址块 10.0.11.64/26 表示的 IP 地址集合是：10.0.11.64(IP 地址低 8 位为 01 000000)～10.0.11.127(IP 地址低 8 位为 01 111111)，和 CIDR 地址块 10.0.11.0/26 聚合后的 IP 地址集合为 10.0.11.0(IP 地址低 8 位为 0 0000000)～10.0.11.127(IP 地址低 8 位为 0 1111111)，实际上就是 CIDR 地址块10.0.11.0/25。

4. 最长前缀匹配

图 5.12 中，假定 IP 分组的目的 IP 地址为 192.1.2.150，路由器必须找出通往 IP 地址为 192.1.2.150 的目的终端的传输路径，由于路由器 R1 在用 IP 地址 192.1.2.150 检索路由表时，判定该 IP 地址包含在由目的网络字段值 192.1.2.0/24 表示的 CIDR 地址块中(根据 24 位网络前缀计算出的目的 IP 地址 192.1.2.150 对应的网络前缀地址＝192.1.2.0)，该 IP 分组被转发给由该路由项指定的下一跳路由器(路由器 R2)，同样，路由器 R2 用 IP 地址 192.1.2.150 检索路由表时，判定该 IP 地址包含在由目的网络字段值 192.1.2.144/28 表示的 CIDR 地址块中(根据 28 位网络前缀计算出的目的 IP 地址 192.1.2.150 对应的网络前缀地址＝192.1.2.144)，该 IP 分组被转发给为第 6 组配置的网络。

图 5.12 中第 6 组对应的网络为了达到既提高访问外部网络的速度，又不改变自己的配置和访问其他组终端的速度的目的，采用同时连接路由器 R1 和路由器 R2 的方式，如图 5.13 所示。这种情况下，路由器 R1 中的路由项变为两项，分别指向路由器 R1 和第 6 组对应的网络。当路由器 R1 接收到目的 IP 地址为 192.1.2.150 的 IP 分组时，发现该 IP 地址与目的网络字段值 192.1.2.0/24 和 192.1.2.144/28 都匹配，路由器 R1 应该如何转发该 IP 分组？显然，路由器 R1 应该直接将该 IP 分组转发给第 6 组对应的网络，这也是将第 6 组对应的网络直接连接路由器 R1 的原因。路由器 R1 用最长前缀匹配来确定传输路径的优先级。最长前缀匹配是指如果有多个目的网络字段值和某个 IP 地址匹配，则选择网络前缀最长的目的网络字段值作为最终匹配结果。在路由器 R1 的路由项中目的网络字段值 192.1.2.0/24 的网络前缀是 24 位，而目的网络字段值 192.1.2.144/28 的网络前缀为 28 位，选择目的网络字段值 192.1.2.144/28 作为最终匹配结果。因此，将 IP 分组直接转发给第 6 组对应的网络。

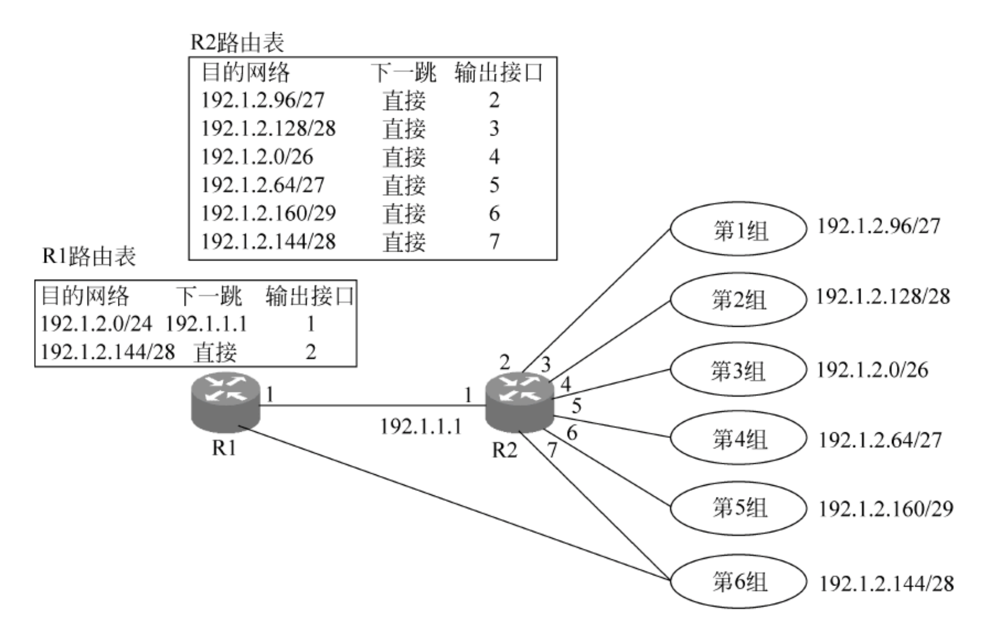


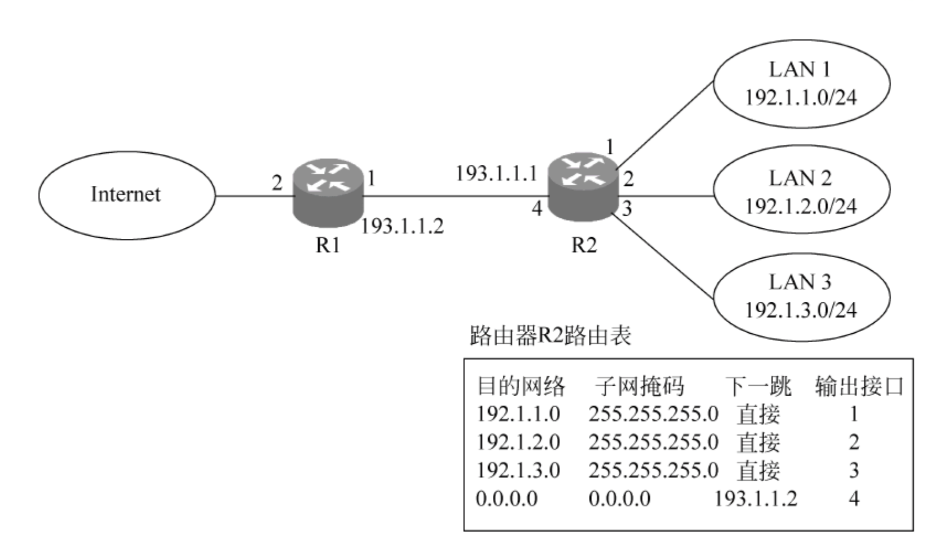
图 5.13 最长前缀匹配过程

5. 默认路由项

如果某个 IP 分组的目的 IP 地址和路由表中所有路由项的目的网络字段值均不匹配，或者丢弃该 IP 分组，或者选择默认路由项指定的传输路径。默认路由项的目的网络字段值为 0.0.0.0，对应的子网掩码为 0.0.0.0，表明所有 IP 地址都和默认网络地址0.0.0.0/0.0.0.0匹配。当通往多个网络的传输路径具有相同的下一跳时，可用一项默认路由项指明通往这些网络的传输路径。图 5.14 所示互连网络中，内部网络通过路由器 R1 连接 Internet，由于 Internet 由无数个网络组成，如果在路由器 R2 的路由表中详细列出 Internet 中所有网络对应的路由项，路由项数目将十分庞大。根据图 5.14 所示互连网络结构，路由器 R2 通往 Internet 的传输路径有着唯一的下一跳：路由器 R1，因此，除了用于指明通往内部网络的传输路径的路由项外，可用一项默认路由项指明通往 Internet 的传输路径。如果某个 IP 地址和 3 个内部网络的网络地址都不匹配，意味着该 IP 地址标识的目的终端位于 Internet，选择通往 Internet 的传输路径转发目的地址为该 IP 地址的 IP 分组。

6. 无分类编址与子网和超网

子网地址是无分类编址出现前的编址方式，一个单位需要分成若干组，每一个组出于安全考虑需要构建独立网络，但是每一个组的终端数较少，单独使用一个网络地址（如 C 类地址）会造成浪费，而且可能整个单位只分配到一个网络地址（如 C 类地址）。这种情况下，需要将单个网络地址分解为若干个子网地址。假如单位分配的 C 类地址是 192.1.1.0，需要将该 C 类地址均匀分配给 6 个子网，这样，每一个子网的主机号字段位数为 8−3=5，3 是子网号的位数，因为子网号的位数是满足 $2^n-2\geqslant$子网数的最小 n，之所以减 2，是因为当时规定子网字段值全 0 和全 1 都不能作为子网号。这种情况下，每一个子网对应的子网掩码为 255.255.255.224，6 个子网对应的子网地址分别是 192.1.1.32/255.255.255.224、



目的网络	子网掩码	下一跳	输出接口
192.1.1.0	255.255.255.0	直接	1
192.1.2.0	255.255.255.0	直接	2
192.1.3.0	255.255.255.0	直接	3
0.0.0.0	0.0.0.0	193.1.1.2	4

图 5.14 默认路由项功能

192.1.1.64/255.255.255.224、192.1.1.96/255.255.255.224、192.1.1.128/255.255.255.224、192.1.1.160/255.255.255.224、192.1.1.192/255.255.255.224。在出现无分类编址技术后，原来的地址类型已不复存在，32 位 IP 地址统一由网络前缀和主机号组成，不存在将某类网络地址划分为多个子网地址的问题。

出现无分类编址后，网络的主机号位数可以是 $N(N\geqslant 2)$，如果某个网络的终端数大于 $2^{10}-2$，小于等于 $2^{11}-2$，需要 11 位主机号，网络前缀位数为 $32-11=21$，可以选择网络地址 192.1.16.0/21。这个网络地址包含的 IP 地址范围恰好是 8 个 C 类地址 192.1.16.0～192.1.23.0 的 IP 地址的集合，将这样的网络地址称为超网地址。有时也将路由项聚合后生成的 CIDR 地址块称为超网地址，用于说明是多个网络的网络地址聚合后生成的 CIDR 地址块。无分类编址中任何由 IP 地址和网络前缀位数确定的 IP 地址集合的确切称呼是 CIDR 地址块，只是该 CIDR 地址块可以是单个网络的 IP 地址集合，也可以是多个网络的 IP 地址集合。

7. 举例

【例 5.3】 路由器的路由表如表 5.1 所示，回答以下问题：

① 假定路由器接收到目的 IP 地址为 142.150.71.132 的 IP 分组，给出路由器为该 IP 分组选择的下一跳，并说明理由。

② 在路由表中增加一项路由项，该路由项的作用是仅仅将目的 IP 地址为 142.150.71.132 的 IP 分组转发给下一跳：A，对目的 IP 地址为其他地址的 IP 分组的转发操作没有任何影响。

③ 在路由表中增加一项路由项，使所有目的 IP 地址中和路由表中已有路由项均不匹配的 IP 分组被转发给下一跳：E。

④ 将 142.150.64.0/24 划分为 4 个规模尽可能大的等长子网，给出每一个子网的可分配地址范围和子网掩码。

表 5.1 路由表

目的网络及前缀	下一跳
142.150.64.0/24	A
142.150.71.128/28	B
142.150.71.128/30	C
142.150.0.0/16	D

【解析】 ①根据路由项的目的网络字段值给出的网络前缀位数确定目的 IP 地址对应的网络前缀地址，目的 IP 地址 142.150.71.132 根据网络前缀位数 24、28、30 和 16 分别确定的网络前缀地址为 142.150.71.0、142.150.71.128、142.150.71.132 和 142.150.0.0。可以看出，目的 IP 地址 142.150.71.132 对应的网络前缀地址与路由项中目的网络字段值分别为 142.150.71.128/28 和 142.150.0.0/16 的路由项匹配，根据最长前缀匹配规则，选择路由项＜142.150.71.128/28，B＞作为最终匹配的路由项，确定下一跳为 B。这一问的关键是根据网络前缀长度求出目的 IP 地址对应的网络前缀地址，假定网络前缀长度为 N，求出目的 IP 地址对应的网络前缀地址的过程如下，用最高 N 位为 1 的 32 位二进制数和 32 位目的 IP 地址进行"与"操作，得出的结果就是该目的 IP 地址对应的网络前缀地址。

② 要求路由项的目的网络字段值唯一匹配目的 IP 地址 142.150.71.132，确定增加的路由项为＜142.150.71.132/32，A＞。

③ 增加默认路由项＜0.0.0.0/0，E＞。

④ 把原来单个网络地址空间分成 4 个等长的网络地址空间，需要把原来的主机号位数减少 2 位，由 8 位变为 6 位，并用这减下的 2 位标识 4 个不同的子网，这样，每一个网络的网络号位数由 24 位变为 26 位，将 142.150.64.0/24 展开，用 26 位网络号重新划分子网后，可以得出如下 4 个子网的地址范围。

二进制地址	子网
10001110 10010110 01000000 **00** 000000 10001110 10010110 01000000 **00** 111111	子网 0，地址范围：142.150.64.0～142.150.64.63； 可分配地址范围：142.150.64.1～142.150.64.62； 子网掩码：255.255.255.192。
10001110 10010110 01000000 **01** 000000 10001110 10010110 01000000 **01** 111111	子网 1，地址范围：142.150.64.64～142.150.64.127； 可分配地址范围：142.150.64.65～142.150.64.126； 子网掩码：255.255.255.192。
10001110 10010110 01000000 **10** 000000 10001110 10010110 01000000 **10** 111111	子网 2，地址范围：142.150.64.128～142.150.64.191； 可分配地址范围：142.150.64.129～142.150.64.190； 子网掩码：255.255.255.192。
10001110 10010110 01000000 **11** 000000 10001110 10010110 01000000 **11** 111111	子网 3，地址范围：142.150.64.192～142.150.64.255； 可分配地址范围：142.150.64.193～142.150.64.254； 子网掩码：255.255.255.192。

5.2.4 IP 分组格式

1. 首部字段

IP 分组由首部与数据两部分组成。首部由 20 个字节的固定项和可变长度的可选项组

成。IP 分组首部格式如图 5.15 所示。



图 5.15　IP 分组首部格式

下面介绍 IP 分组首部各字段的意义。

版本：4b 版本字段给出 IP 分组所属 IP 协议的版本。由于每一个 IP 分组都含有版本字段，就允许存在几个月，甚至几年时间的升级过渡期，在这段时间内，不同版本的 IP 协议可同时在一个网络内运行。目前存在两种版本的 IP 协议：IPv4 和 IPv6，其版本号分别为 4 和 6。

首部长度：4b 首部长度字段以 32 位字为单位给出 IP 首部的实际长度。由于首部的长度不是固定的，需要用首部长度字段给出 IP 首部长度。字段最小值为 5，用于没有可选项的情况，由于 IP 首部长度的基本单位是 4 个字节，意味着首部固定部分长度为 20 字节。最大值为 15，这就将首部长度限制在 60 个字节内，意味着可选项长度不能超过 40 个字节。

服务类型：8b 服务类型字段允许终端告诉网络它希望得到的服务，可以通过服务类型字段指定 IP 分组的速度要求、可靠性要求及各种要求的组合。不同应用有不同的性能要求，对于数字化语音，快速到达比正确到达更重要。对于文件传送，无错传送比快速传送更为重要。服务类型字段从左到右包括三位优先级位，三位标志位 D、T、R 和目前没有使用的二位。三位优先级位表示从 0(普通报文)到 7(网络控制报文)8 级分组优先级，优先级高的 IP 分组优先得到服务。三位标志位允许终端指定最希望得到的服务。允许指定的服务是 D：时延，T：吞吐率，R：可靠性。D=1 表示该 IP 分组要求特别短的时延。T=1 表示该 IP 分组要求特别高的吞吐率。R=1 表示要求该 IP 分组尽可能不被损坏或丢弃。这些标志位可以帮助路由器选择对应的传输路径。实际上，早先的路由器一般都不考虑这些标志位，目前为支持多媒体应用，路由器开始支持服务分类(CoS)。

总长度：16b 总长度字段以字节为单位给出包括首部和数据的 IP 分组的长度，最大长度值为 65535 字节。根据目前存在的传输网络状况，这个值是绰绰有余了。

标识：16b 标识字段告诉目的终端，那些数据片是属于同一 IP 分组的，属于同一 IP 分组的数据片具有相同标识字段值。发送端维持一个计数器，每发送一个 IP 分组，计数器加 1，计数器的值就作为 IP 分组的标识字段值。

标志：目前定义了 2 位标志位 DF 和 MF 位。DF 位置 1 要求不能对 IP 分组分片，它命令路由器不要把 IP 分组分片成数据片，因为目的终端没有能力把分片后的数据片重新装配成 IP 分组，例如：计算机引导时，ROM 要求把存储映像作为单个 IP 分组送给它。一旦 IP 分组中的 DF 位置 1，表明该 IP 分组只能作为单个数据片传送，这就要求路由器即使选择一条并不是最佳的路由，也要避开只能传输长度很短的 IP 分组的传输网络。要求所有网络至

少能传输小于576字节的IP分组。MF位置0表示是若干数据片中最后一个数据片，除最后一个数据片外，IP分组分片后所生成的所有其他数据片都必须将MF位设置为1，MF位的作用是使接收终端知道某个IP分组分片后所生成的所有数据片是否已全部接收到。

片偏移：13b片偏移字段以8个字节为单位给出该数据片在分片前的原始数据中的起始位置。因此，除最后数据片以外的所有其他数据片，它们的长度必须是8字节的整数倍。由于该字段有13位，由此可推出IP分组的最大长度为$2^{13}\times8=65536$字节。

生存时间：字段长度8b，此字段是用于限制IP分组存在时间的一个计数器，假定该计数器以秒为单位计数，IP分组允许存在的最长时间为255秒。目前，该字段只是作为最大跳数使用，IP分组每经过一跳路由器，该字段值减1，当值减为0时，丢弃该IP分组并发送一个警告消息给源终端。设置该字段的目的是避免IP分组因为路由器的路由表被破坏而使IP分组在网络上无休止地漂荡。

协议：字段长度16b，当网络层把完整的IP分组装配好以后，它需要知道如何处理该IP分组。协议类型字段告诉网络层把该IP分组提交给哪一个进程处理。TCP进程和UDP进程是最有可能处理该IP分组的进程。

首部检验和：字段长度16b，将IP分组首部以16位为单位分段，然后根据反码运算规则对各段进行累加，将累加结果求反作为首部检验和。首部检验和的作用是检测出首部传输过程中发生的错误。首部检验和经过每一跳路由器都必须重新计算一次，因为每经过一跳至少改变了一个首部字段值(生存时间字段)。

源地址和目的地址：字段长度32b，该字段给出了源终端的网络号和主机号及目的终端的网络号和主机号。

可选项：设计该字段的目的主要是：①允许以后协议版本提供原始设计中遗漏的信息；②允许经验丰富的人试验一些新的想法；③避免在报文首部中固定分配一些并不常用的信息字段。可选项长度可变。目前，定义了5种可选项，如表5.2所示，需要强调的是，并不是所有路由器都支持这5种可选项。

表5.2 IP可选项

可 选 项	描 述
保密	指定IP分组如何保密
严格的源站选路	给出用于传输IP分组的完整路由
不严格的源站选路	给出不允许遗漏的一些路由器列表
记录路由	每一个经过的路由器将它的IP地址添加到IP分组中
时间戳	每一个经过的路由器将它的IP地址和时间戳添加到IP分组中

保密：该选项给出如何保密IP分组，与军事应用有关的路由器可以用该选项来避开某些认为不安全的国家或地区。实际上，所有路由器都忽略该选项。

严格的源站选路：该选项给出源终端至目的终端传输路径完整的IP地址列表，IP分组必须严格遵循给出的传输路径。系统管理员可以用这种功能在路由器路由表损坏的情况下发送紧急IP分组，或者用于发送测量时间参数的IP分组。

不严格的源站选路：该选项要求IP分组一定要经过列表中指定的路由器，并按指定的顺序经过。但允许通过传输路径上别的路由器。通过用该选项指定少数几个路由器来强迫

IP 分组经过某一特殊传输路径。例如：强迫从伦敦到悉尼的 IP 分组经过美国西部而不是东部时，该选项可指定 IP 分组必须经过纽约、洛杉矶、檀香山的路由器。当出于某种政治或经济考虑，需要 IP 分组经过或避开某些地区或国家时，可用该选项。

记录路由：该选项要求所有经过的路由器把它们的 IP 地址添加到该选项字段中，通过记录路由，可以帮助系统管理员查出路由算法中的一些问题。由于 ARPA 网刚建立时，IP 分组经过的路由器最多不超过 9 个，因此用 40 个字节记录经过的路由器已经很充足了，但对现在的 Internet 来说，用 40 个字节记录经过的路由器是远远不够的。

时间戳：该选项基本上与记录路由选项一样，不同的是，除记录 32 位的 IP 地址外，还记录 32 位的时间戳。该选项也主要用于诊断路由算法发生的错误。

IP 分组首部的可选项有很强的了解、管理网络的功能，常常被用来作为侦察网络的工具，为了网络的安全性，路由器需要关闭一些可选项的支持功能。

2. 分片

传输网络链路层帧净荷字段（也称载荷字段）允许的最大长度称为最大传输单元(Maximum Transfer Unit，MTU)，如以太网的 MTU 为 1500B，如果 IP 分组长度超过传输该 IP 分组的传输网络的 MTU，必须将 IP 分组分片，分片过程是将 IP 分组净荷字段中的数据分片为多个数据片，除了最后一个数据片，其他数据片的长度必须是 8B 的整数倍。每一个数据片加上 IP 首部构成 IP 分组，必须保证分片后的数据片长度和 IP 首部长度之和小于传输网络的 MTU。通常情况下，除最后一个数据片，其他数据片长度的分配原则是：必须是 8 的倍数，且加上 IP 首部后尽量接近 MTU。为了标识这些由分片同一个 IP 分组净荷字段中的数据产生的 IP 分组序列，这些 IP 分组必须具有相同的标识字段值，为了在目的端将这些 IP 分组中净荷字段包含的数据片重新还原为原始数据，这些 IP 分组中的每一个 IP 分组必须在片偏移字段中给出该 IP 分组包含的数据片在原始数据中的起始位置，为了让目的端确定所有数据片对应的 IP 分组均已到达，必须标志最后一个数据片对应的 IP 分组。分片过程如图 5.16 所示，4000B 数据被分成 3 个数据片，长度分别是 1480B、1480B 和 1040B。

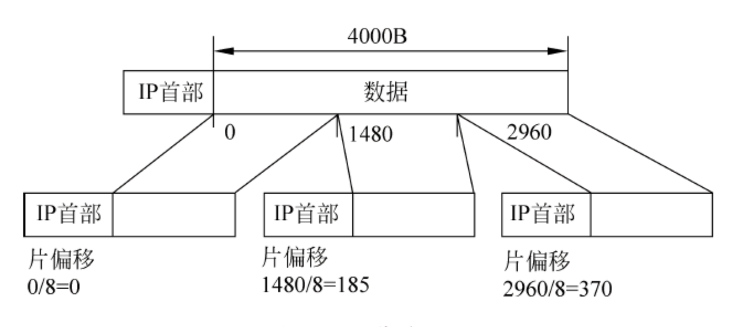


图 5.16 分片过程

【例 5.4】 终端 A 和终端 B 之间传输路径由网络 1、网络 2 和网络 3 组成，其中网络 1 的 MTU＝1500B，网络 2 的 MTU＝800B，网络 3 的 MTU＝420B，假定终端 A 传输给终端 B 的数据的长度为 1440B，给出终端 A 及传输路径经过的各个路由器分片数据的过程。

【解析】 终端 A 及传输路径经过的路由器分片数据的过程如图 5.17 所示。终端 A 生成的 IP 分组的总长度为 1460B(包括 20B 首部和 1440B 净荷)，由于终端 A 连接路由器 R1

的链路的 MTU＝1500B，终端 A 可以直接将总长度为 1460B 的 IP 分组传输给路由器 R1。当路由器 R1 向路由器 R2 传输该 IP 分组时，发现输出链路的 MTU＝800B，需要对 IP 分组进行分片操作。路由器 R1 将 IP 分组的净荷分成 2 个数据片，2 个数据片的长度分别为 776B 和 664B，加上 20B 的 IP 首部后，分别构成 2 个总长度分别为 796B（20B 首部＋776B 净荷）和 684B 的 IP 分组。这 2 个 IP 分组的标识符字段值相同，后一个 IP 分组的片偏移＝776/8＝97。同样，当路由器 R2 向终端 B 传输这 2 个 IP 分组时，发现输出链路的 MTU＝420B，路由器 R2 需要再一次对这 2 个 IP 分组进行分片操作，776B 的数据片被分片成长度分别为 400B 和 376B 的 2 个数据片，同样，664B 数据片被分片成长度分别为 400B 和 264B 的 2 个数据片，这 4 个数据片加上 IP 首部后构成 4 个 IP 分组，原来 M 标志位为 1 的 IP 分组分片后生成的 IP 分组序列的 M 标志位都为 1。原来 M 标志位为 0 的 IP 分组分片后生成的 IP 分组序列，除由最后一个数据片构成的 IP 分组外，其他 IP 分组的 M 标志位也都为 1。这些 IP 分组的标识字段值都相同，图中每一个 IP 分组首部中的片偏移给出净荷中的数据片在原始净荷中的位置。

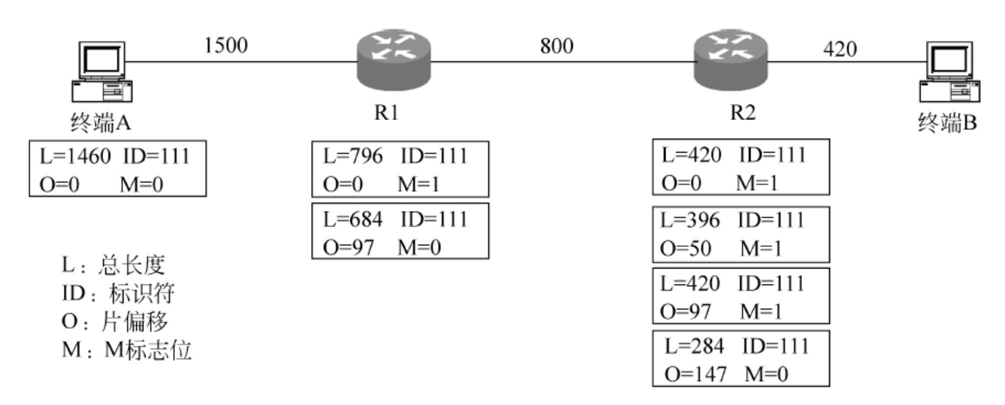


图 5.17 分片数据过程

5.3 路由表和 IP 分组端到端传输过程

5.3.1 路由表建立过程

1. 简单互连网络路由表生成过程

实现 IP 分组逐跳传输的关键是每一个路由器建立路由表，路由表中每一项路由项给出通往某个目的终端的传输路径。如果某个路由器通往一组目的终端的传输路径有着相同的下一跳，用一项路由项给出通往一组目的终端的传输路径。路由项的通用格式为＜目的网络字段值，下一跳 IP 地址＞，目的网络字段值给出表示一组目的终端的 IP 地址集合的网络前缀地址，下一跳地址值是下一跳连接互连当前跳和下一跳网络的接口的 IP 地址。

路由器每一个接口连接一个网络，该网络的网络地址由配置给该路由器接口的 IP 地址和子网掩码决定，如果某个路由器接口配置 IP 地址和子网掩码 192.1.1.254/255.255.255.0，意味着该接口连接的网络的网络地址为 192.1.1.0/24。连接在该网络上的所有终

端的 IP 地址必须属于网络地址 192.1.1.0/24。对于图 5.18 所示网络结构，路由器 R1 接口 1 连接的网络 LAN 1 的网络地址为 192.1.1.0/24，路由器 R3 接口 1 连接的网络 LAN 2 的网络地址为 192.1.2.0/24，如果要求实现 LAN 1 中终端和 LAN 2 中终端之间的相互通信，必须建立 LAN 1 与 LAN 2 之间的双向传输路径。对于路由器 R1，LAN 1 是接口 1 直接连接的网络，通往 LAN 2 的传输路径上的下一跳是路由器 R2，路由器 R2 连接路由器 R1 的接口的 IP 地址是 192.2.1.2，因此，路由器 R1 分别生成用于指明通往网络 192.1.1.0/24 和网络 192.1.2.0/24 的传输路径的两项路由项＜192.1.1.0/24，直接＞和＜192.1.2.0/24，192.2.1.2＞。对于路由器 R2，通往 LAN 1 的传输路径上的下一跳是路由器 R1，路由器 R1 连接路由器 R2 的接口的 IP 地址是 192.2.1.1，通往 LAN 2 的传输路径上的下一跳是路由器 R3，路由器 R3 连接路由器 R2 的接口的 IP 地址是 192.2.2.2，因此，路由器 R2 分别生成用于指明通往网络 192.1.1.0/24 和网络 192.1.2.0/24 的传输路径的两项路由项＜192.1.1.0/24，192.2.1.1＞和＜192.1.2.0/24，192.2.2.2＞。对于路由器 R3，LAN 2 是接口 1 直接连接的网络，通往 LAN 1 的传输路径上的下一跳是路由器 R2，路由器 R2 连接路由器 R3 的接口的 IP 地址是 192.2.2.1，因此，路由器 R3 分别生成用于指明通往网络 192.1.1.0/24 和网络 192.1.2.0/24 的传输路径的两项路由项＜192.1.1.0/24，192.2.2.1＞和＜192.1.2.0/24，直接＞。各个路由器最终生成的用于指明通往网络 LAN 1 和 LAN 2 的传输路径的路由项如图 5.18 所示。

路由器通过路由表确定通往网络 LAN 1 和 LAN 2 的传输路径，连接 LAN 1 上的终端通过默认网关地址确定通往 LAN 2 的传输路径，该默认网关地址是路由器 R1 连接 LAN 1 的接口(接口 1)的 IP 地址。终端的默认网关地址给出该终端通往其他网络的传输路径上的第一跳路由器地址，该终端传输给其他网络上终端的 IP 分组，首先传输给与其连接在同一个网络上的、用默认网关地址指定的第一跳路由器。如果有多个路由器连接在终端所连接的网络上，可以选择其中一个路由器作为第一跳路由器，用该路由器连接终端所在网络的接口的 IP 地址作为终端的默认网关地址。

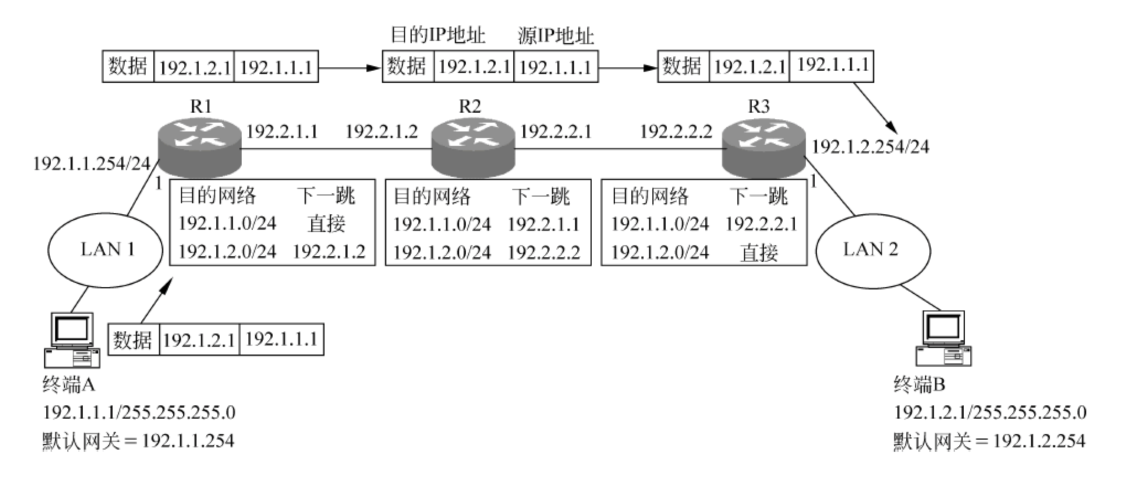


图 5.18　IP 分组传输过程

2. 复杂互连网络路由表生成过程

复杂互连网络结构如图 5.19 所示。假定路由器 R5 要为 4 个网络(网络地址分别为

192.1.1.0/24,192.1.2.0/24,192.1.3.0/24,192.1.4.0/24)选择最少跳数的传输路径,即经过的路由器数目最少的传输路径(这种传输路径也称为最短路径)。管理员可以通过分析图 5.19 所示的互连网络拓扑结构,确定路由器 R5 通往这 4 个网络的最短路径,如图 5.19 中箭头指出的传输路径,并因此给出表 5.3 所示的路由表,其中目的网络字段给出需要到达的网络的网络地址,距离字段给出路由器 R5 到达目的网络的最短路径所经过的路由器数目(含路由器 R5),下一跳字段给出路由器 R5 通往目的网络的最短路径上的下一跳路由器,如果目的网络和路由器 R5 直接相连,下一跳字段中用直接表示。

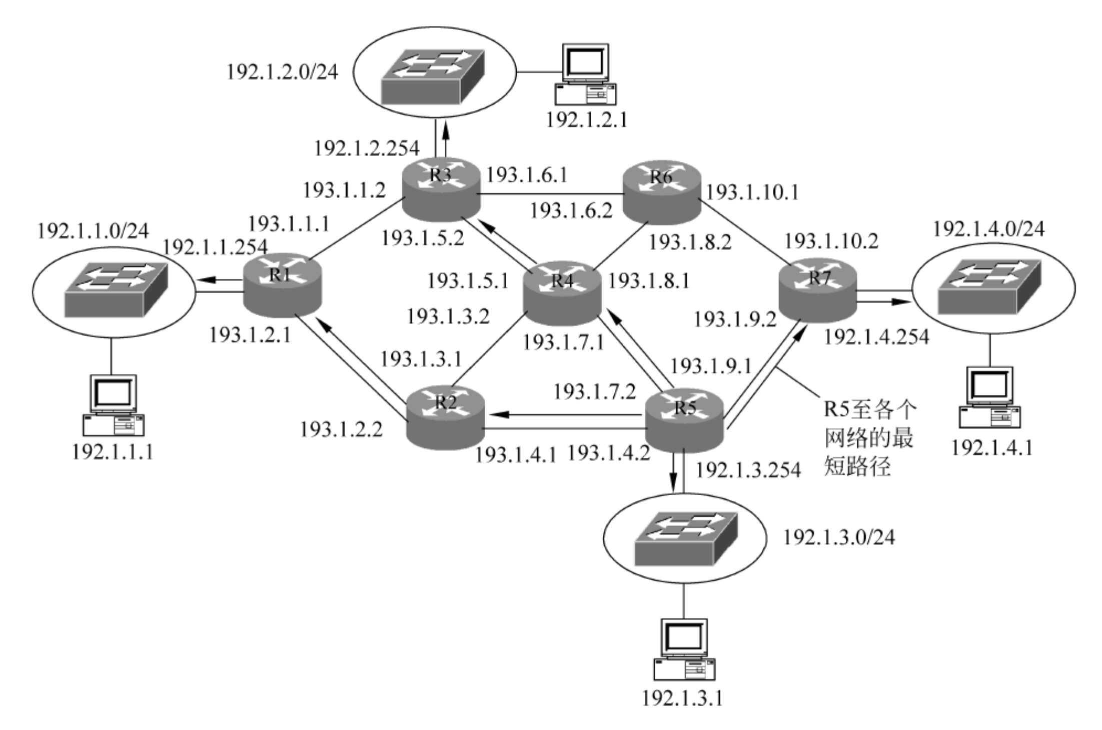


图 5.19 人工配置静态路由表过程

表 5.3 路由器 R5 配置的静态路由表

目的网络	距离	下一跳路由器
192.1.1.0/24	3	193.1.4.1
192.1.2.0/24	3	193.1.7.1
192.1.3.0/24	1	直接
192.1.4.0/24	2	193.1.9.2

3. 例题解析

【例 5.5】 互连网络结构如图 5.20 所示,路由表每一项路由项包含字段<目的网络,子网掩码,下一跳,输出接口>,回答下列问题:

① 将 IP 地址空间 202.115.1.0/24 划分为两个子网,分别分配给 LAN 1 和 LAN 2,每个子网分配的 IP 地址数不少于 120,给出子网划分结果,说明理由并给出必要的计算过程;

② 给出路由器 R1 的路由表,包含用于指明通往图 5.20 中所有网络和服务器的传输路径的路由项;

③ 给出路由器 R2 路由表中用于指明通往 LAN 1 和 LAN 2 的传输路径的路由项。

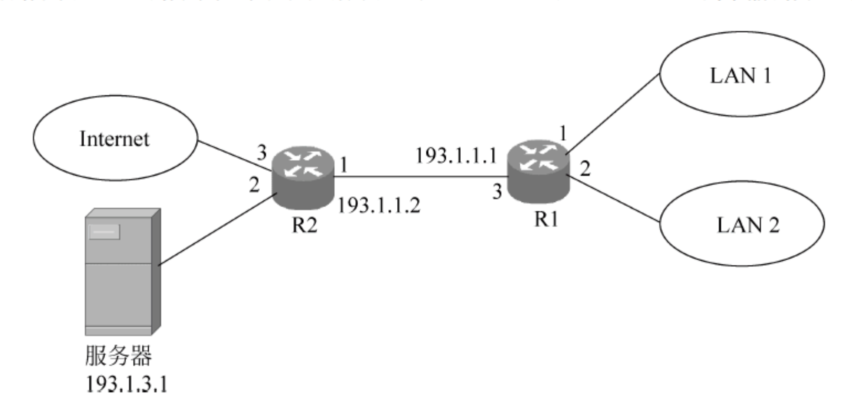


图 5.20　互连网络结构

【解析】 ① 由于要求每一个子网的 IP 地址数不小于 120，因此，每一个子网至少用 7 位二进制数作为主机号，这样，可以把给定的 IP 地址空间分成两个等长的网络地址空间，把原来的主机号位数减少 1 位，由 8 位变为 7 位，并用这减下的 1 位标识两个不同的子网，这样，每一个网络地址空间中的网络号位数由 24 位变为 25 位，将 202.115.1.0/24 展开，用 25 位网络号重新划分子网后，可以得出如下两个子网的 IP 地址范围。

11001010 01110011 00000001 **0** 0000000
10001110 10010110 01000000 **0** 1111111

子网 0，地址范围：202.115.1.0～202.115.1.127；
网络地址：202.115.1.0；
子网掩码：255.255.255.128。

11001010 01110011 00000001 **1** 0000000
10001110 10010110 01000000 **1** 1111111

子网 1，地址范围：202.115.1.128～202.115.1.255；
网络地址：202.115.1.128；
子网掩码：255.255.255.128。

② 路由器 R1 的路由表如表 5.4 所示。由于两个子网和路由器 R1 直接相连，表中用于指明通往两个子网的传输路径的路由项的下一跳字段值为直接。用于指明通往服务器的传输路径的路由项中的目的网络和子网掩码字段值必须唯一匹配该服务器的 IP 地址。因此，用 32 位全 1 的子网掩码，表示目的网络只包含单个 IP 地址：193.1.3.1。由于 Internet 是由无数个网络组成的，因此，只能以默认路由项指明通往 Internet 的传输路径。由于路由器 R1 至服务器和 Internet 传输路径上的下一跳是路由器 R2，且路由器 R2 接口 1 和路由器 R1 接口 3 连接在同一个网络上，用路由器 R2 接口 1 的 IP 地址作为这些路由项的下一跳 IP 地址。

表 5.4　路由器 R1 路由表

目的网络	子网掩码	下一跳	输出接口
202.115.1.0	255.255.255.128	直接	1
202.115.1.128	255.255.255.128	直接	2
193.1.3.1	255.255.255.255	193.1.1.2	3
0.0.0.0	0.0.0.0	193.1.1.2	3

③ 路由器 R2 应该有两项路由项分别用于指明通往 LAN 1 和 LAN 2 的传输路径，但这两项路由项有着相同的输出接口和下一跳，这样的路由项可以尝试聚合为一项路由项，前提是聚合后的目的网络和子网掩码字段确定的 IP 地址空间等于聚合前两项路由项所包含的 IP 地址空间，这里，(202.115.1.0～202.115.1.127)＋(202.115.1.128～202.115.1.255)＝(202.115.1.0～202.115.1.255)，因此，可以完成图 5.21 所示的合并过程。

目的网络	子网掩码	下一跳	接口
202.115.1.0	255.255.255.128	193.1.1.1	1
202.115.1.128	255.255.255.128	193.1.1.1	1

目的网络	子网掩码	下一跳	接口
202.115.1.0	255.255.255.0	193.1.1.1	1

图 5.21　路由项合并过程

5.3.2　IP 分组端到端传输过程

为了深刻理解路由表的作用，详细讨论图 5.18 中终端 A 至终端 B 的 IP 分组传输过程。

1. 确定源终端和目的终端是否在同一个网络

终端 A 向终端 B 传输数据前，必须先获取终端 B 的 IP 地址，然后将数据封装成以终端 A 的 IP 地址为源 IP 地址，以终端 B 的 IP 地址为目的 IP 地址的 IP 分组，在进行 IP 分组传输前，先确定终端 B 是否和终端 A 位于同一个网络，步骤如下：

(1) 终端 A 根据自己的 IP 地址和子网掩码，求出网络地址：192.1.1.0。

(2) 终端 A 根据终端 B 的 IP 地址和自己的子网掩码，求出终端 B 的网络地址：192.1.2.0。

(3) 如果两个网络地址相同，说明终端 A 和终端 B 位于同一个网络，终端 A 至终端 B 的 IP 分组传输过程无须经过路由器。

(4) 如果两个网络地址不相同，说明终端 A 和终端 B 位于不同的网络，终端 A 将沿着由路由器构成的 IP 分组传输路径，逐跳转发 IP 分组。

2. 根据默认网关地址找到第一跳路由器

一旦确定终端 B 和终端 A 不在同一个网络，终端 A 将 IP 分组转发给终端 A 至终端 B 传输路径上的第一跳路由器，该路由器的 IP 地址通过配置的默认网关地址获得，这里的默认网关实际上是默认路由器，因此，也将默认网关地址称为默认路由器地址。如果连接终端 A 和第一跳路由器的网络是以太网，必须将 IP 分组封装成以终端 A 的 MAC 地址为源 MAC 地址，以第一跳路由器连接以太网的接口的 MAC 地址为目的 MAC 地址的 MAC 帧，然后将 MAC 帧通过以太网传输给第一跳路由器。

3. 路由器逐跳转发

IP 分组到达路由器 R1 后，路由器 R1 根据最长前缀匹配算法用 IP 分组的目的 IP 地址匹配路由表中的所有路由项，如果找到匹配的路由项，将 IP 分组转发给路由项指定的下一跳。路由器 R1 的路由表中，只有路由项＜192.1.2.0/24，192.2.1.2＞和 IP 分组的目的 IP

地址匹配，IP 分组被转发给 IP 地址为 192.2.1.2 的下一跳路由器。传输路径上的路由器依次逐跳转发，IP 分组到达传输路径上最后一跳路由器 R3。

4. 直接交付

路由器 R3 中和 IP 分组目的 IP 地址匹配的路由项是＜192.1.2.0/24，直接＞，表明该路由器和终端 B 之间不再有其他路由器，即终端 B 和该路由器的其中一个接口连接在同一个网络上，路由器通过该网络将 IP 分组直接传输给终端 B。如果连接路由器 R3 和终端 B 的网络是以太网，必须将 IP 分组封装成以路由器 R3 连接以太网的接口的 MAC 地址为源 MAC 地址、以终端 B 的 MAC 地址为目的 MAC 地址的 MAC 帧，然后将 MAC 帧通过以太网传输给终端 B。

从上述讨论的 IP 分组端到端传输过程可以得出以下实现 IP 分组端到端传输的基本思路：

(1) 建立一条以源终端为始点，以目的终端为终点，中间由若干路由器组成的 IP 分组端到端传输路径，IP 分组沿着端到端传输路径逐跳转发。源终端通过配置的默认网关地址获得第一跳路由器的 IP 地址，中间路由器根据路由表和 IP 分组的目的 IP 地址确定下一跳路由器地址；

(2) 在获取下一跳路由器的 IP 地址后，通过 IP over X 技术，实现 IP 分组当前跳至下一跳的传输过程，X 是连接当前跳和下一跳的传输网络，如以太网。

建立端到端传输路径的关键是每一个路由器建立路由表，路由表中每一项路由项指出通往特定网络的传输路径上的下一跳路由器，因此，解决 IP 分组端到端传输的第一步是为互连网络中的每一个路由器建立路由表。

5. 例题解析

【例 5.6】 确定下述主机对是否连接在同一个网络上。

① 主机 1：172.16.5.72/255.255.255.0，主机 2：172.16.5.79/255.255.255.0。

② 主机 1：192.168.19.35/255.255.255.224，主机 2：192.168.19.48/255.255.255.224。

③ 主机 1：10.128.14.14/255.255.255.240，主机 2：10.128.14.19/255.255.255.240。

④ 主机 1：192.168.3.68/255.255.255.248，主机 2：192.168.3.74/255.255.255.248。

【解析】 判定两个主机是否连接在同一个网络上的依据是这两个主机的网络地址是否相同，主机的网络地址是主机的 IP 地址与主机的子网掩码进行“与”操作的结果。

① 由于子网掩码是 24 位 1 和 8 位 0，主机的网络地址取 IP 地址的前 24 位，最后 8 位为 0，因此，主机 1 的网络地址＝172.16.5.0/24，主机 2 的网络地址＝172.16.5.0/24，主机 1 和主机 2 连接在同一个网络上。

② 由于子网掩码是 27 位 1 和 5 位 0，且两个 IP 地址的前 24 位相同，因此，只需计算出 IP 地址最后 1 个字节的前 3 位值。

00100011(35)&&11100000(224)＝00100000(32)

00110000(48)&&11100000(224)＝00100000(32)

由此得出主机 1 和主机 2 的网络地址均是 192.168.19.32/27。主机 1 和主机 2 连接在同一个网络上。

③ 由于子网掩码是28位1和4位0,且两个IP地址的前24位相同,因此,只需计算出IP地址最后1个字节的前4位值。

00001110(14)&&11110000(240)=00000000(0)

00010011(19)&&11110000(240)=00010000(16)

由此得出主机1的网络地址=10.128.14.0/28,主机2的网络地址=10.128.14.16/28。主机1和主机2连接在不同的网络上。

④ 由于子网掩码是29位1和3位0,且两个IP地址的前24位相同,因此,只需计算出IP地址最后1个字节的前5位值。

01000100(68)&&11111000(248)=01000000(64)

01001010(74)&&11111000(248)=01001000(72)

由此得出主机1的网络地址=192.168.3.64/29,主机2的网络地址=192.168.3.72/29。主机1和主机2连接在不同的网络上。

5.3.3 ARP和地址解析过程

1. 地址解析过程

图5.22中,假定终端A和服务器B连接在同一个网络上,即使如此,终端A访问服务器B时所给出的也不会是服务器B的MAC地址,往往是服务器B的域名,经过域名服务器解析后得到的也只能是服务器B的IP地址。根据以太网交换机的工作原理,以太网交换机只能根据MAC帧的目的MAC地址和转发表来转发MAC帧,这就意味着:①不能在以太网上直接传输IP分组,必须将IP分组封装成MAC帧;②在将IP分组封装成MAC帧前,必须先获取连接在同一个网络上的源终端和目的终端的MAC地址。源终端的MAC地址可以直接从安装的网卡中读取,问题是如何根据目的终端的IP地址来获取目的终端的MAC地址。地址解析协议(Address Resolution Protocol,ARP)和地址解析过程就用于实现这一功能,ARP请求帧格式如图5.23所示。

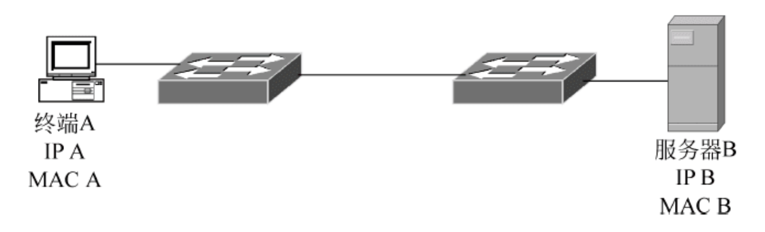


图5.22 以太网内传送IP分组过程

图5.24中,终端A获知了服务器B的IP地址IP B后,广播一个MAC帧,该MAC帧的格式如图5.23所示,它的源MAC地址为终端A的MAC地址MAC A,目的MAC地址为广播地址ff-ff-ff-ff-ff-ff,MAC帧中的数据字段包含终端A的IP地址IP A和MAC地址MAC A,同时,包含服务器B的IP地址IP B,IP B是需要解析的IP地址,称为目标地址。该帧是ARP请求帧,它要求IP地址为IP B的网络终端回复它的MAC地址。

ff-ff-ff-ff-ff-ff	MAC A	类型＝地址解析	数据	FCS
目的 MAC 地址	源 MAC 地址		IP A MAC A IP B ?	

图 5.23 用于地址解析的 MAC 帧

由于该 MAC 帧的目的地址为广播地址，同一网络内的所有终端都能够接收到该 MAC 帧，每一个接收到该 MAC 帧的终端首先检测自己的 ARP 缓冲区，如果 ARP 缓冲区中没有发送终端的 IP 地址和 MAC 地址对，将发送终端的 IP 地址和 MAC 地址对(IP A 和 MAC A)记录在 ARP 缓冲区中，然后比较 MAC 帧中给出的目标 IP 地址是否和自己的 IP 地址相同，如果相同，回复自己的 MAC 地址，整个过程如图 5.24 所示。

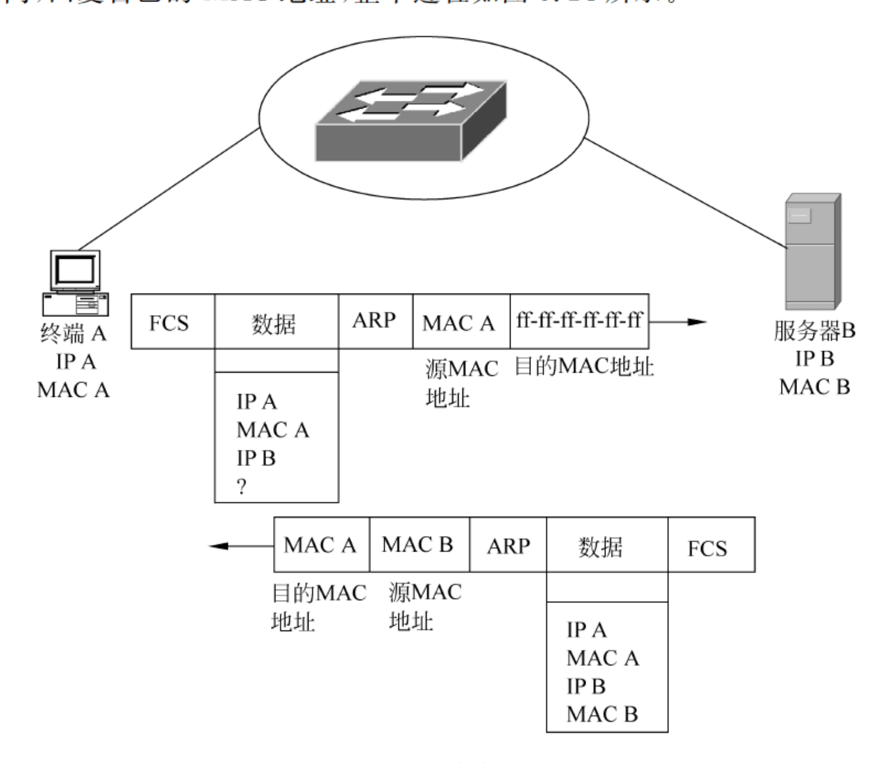


图 5.24 ARP 解析地址过程

ARP 地址解析过程只能发生在连接在同一个以太网上的源终端和目的终端之间，如果源终端和目的终端不在同一个网络内，则 IP 分组需要逐跳转发，源终端必须先将 IP 分组发送给由默认网关地址指定的第一跳路由器，当然，如果连接源终端和第一跳路由器的网络是以太网，源终端通过 ARP 地址解析过程获取第一跳路由器连接以太网的接口的 MAC 地址。同样，如果连接第一跳和下一跳路由器的网络也是以太网，如图 5.25 所示，第一跳路由器也需通过 ARP 地址解析过程获取下一跳路由器连接以太网的接口的 MAC 地址。总之，如果互连当前跳和下一跳的网络是以太网，IP 分组封装成 MAC 帧后才能经过以太网实现当前跳至下一跳的传输过程，在将 IP 分组封装成 MAC 帧前，必须获取下一跳连接以太网

的接口的 MAC 地址，ARP 地址解析过程用于完成根据下一跳连接以太网的接口的 IP 地址求出该接口的 MAC 地址的过程。

2. MAC 帧封装过程

图 5.25 给出了终端 A 传输给终端 B 的 IP 分组经过各个以太网时的封装过程，IP 分组由终端 A 至终端 B 的传输过程中是不变的，但 IP 分组经过互连终端 A 和路由器 R1 的以太网时，封装成以路由器 R1 接口 1 的 MAC 地址 MAC R11 为目的 MAC 地址、以终端 A 的 MAC 地址 MAC A 为源 MAC 地址的 MAC 帧，类型字段 0800 表示净荷是 IP 分组。终端 A 通过解析默认网关地址 192.1.1.254 获得路由器 R1 接口 1 的 MAC 地址。IP 分组经过互连路由器 R1 和路由器 R2 的以太网时，封装成以路由器 R2 接口 1 的 MAC 地址 MAC R21 为目的 MAC 地址、以路由器 R1 接口 2 的 MAC 地址 MAC R12 为源 MAC 地址的 MAC 帧。路由器 R1 通过检索路由表获取路由器 R2 接口 1 的 IP 地址 192.1.3.2，通过解析 IP 地址 192.1.3.2 获得路由器 R2 接口 1 的 MAC 地址。IP 分组经过互连路由器 R2 和终端 B 的以太网时，封装成以终端 B 的 MAC 地址 MAC B 为目的 MAC 地址、以路由器 R2 接口 2 的 MAC 地址 MAC R22 为源 MAC 地址的 MAC 帧。路由器 R2 通过检索路由表得知终端 B 直接连接在接口 2 连接的以太网上，通过解析终端 B 的 IP 地址 192.1.2.1 获得终端 B 的 MAC 地址。

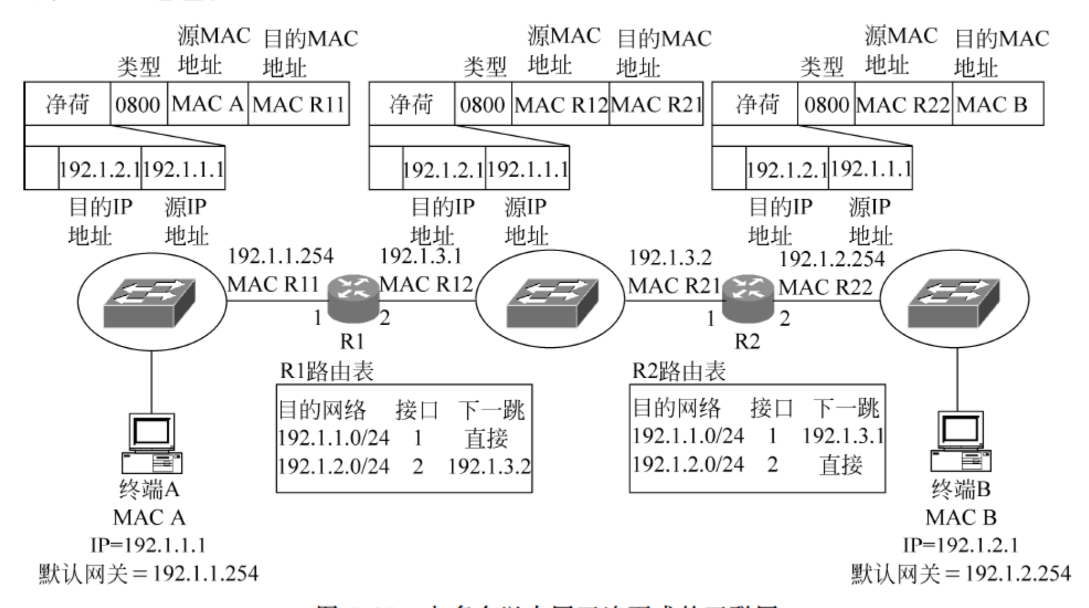


图 5.25 由多个以太网互连而成的互联网

5.4 虚拟路由器冗余协议

5.4.1 容错网络结构

互连网络结构如图 5.26 所示，每一个以太网内部通过链路冗余和生成树协议保证在发生单条链路故障的情况下仍然保持连接在同一以太网上的终端之间的连通性。同时，路由

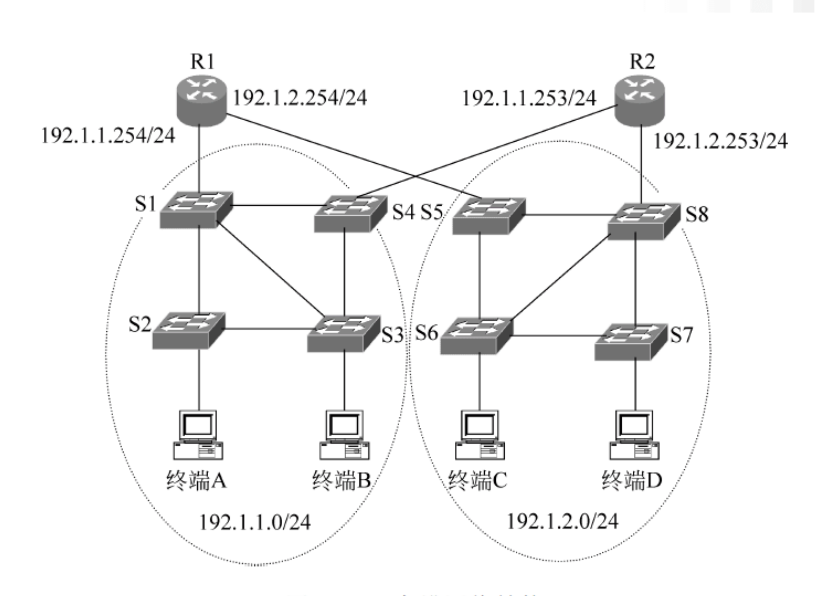


图 5.26 容错网络结构

器 R1 和路由器 R2 分别有接口连接到两个以太网，保证在其中一个路由器发生故障的情况下仍然保持连接在不同以太网上的终端之间的连通性，因此，图 5.26 所示互连网络结构是一种不会因为单点故障导致网络连通性发生问题的容错网络结构。

由于每一个以太网同时连接两个路由器接口，因此，连接在每一个以太网上的终端可以在分配给两个路由器接口的两个 IP 地址中选择一个 IP 地址作为默认网关地址。如终端 A 可以选择 192.1.1.254 或 192.1.1.253 作为默认网关地址，但由于目前终端一般只能配置一个默认网关地址，因此，即使对于图 5.26 所示容错结构，终端在只能配置单个默认网关地址且作为默认网关的路由器失效的情况下，必须通过手工配置新的默认网关地址来保持该终端和其他终端之间的连通性，如果终端 A 配置了默认网关地址 192.1.1.254，一旦路由器 R1 失效，必须通过手工配置方式为终端 A 配置新的默认网关地址 192.1.1.253，否则，终端 A 无法和连接在其他网络上的终端通信。

对于图 5.26 所示的容错网络结构，希望有一种和生成树协议相似的协议，该协议能够根据优先级在多个可以作为默认网关的路由器中选择一个路由器作为其默认网关，一旦该路由器发生故障，能够自动选择另一个路由器作为默认网关，并自动完成两个路由器之间的功能切换。虚拟路由器冗余协议（Virtual Router Redundancy Protocol，VRRP）就是这样一种协议。

5.4.2 VRRP 工作原理

1. VRRP 工作环境

VRRP 工作环境如图 5.27 所示，支持 VRRP 的路由器称为 VRRP 路由器，多个有接口连接在同一个网络上的 VRRP 路由器（如图 5.27 中路由器 R1 和路由器 R2）构成一个虚拟路由器，这些 VRRP 路由器中只有一个 VRRP 路由器是主路由器，其他路由器为备份路由器。VRRP 作用的网络可以是任意支持广播的网络，如以太网、令牌环网和 FDDI，连接在这些网络上的终端和路由器接口有着唯一的 MAC 地址，这里以以太网为例来讨论 VRRP 的

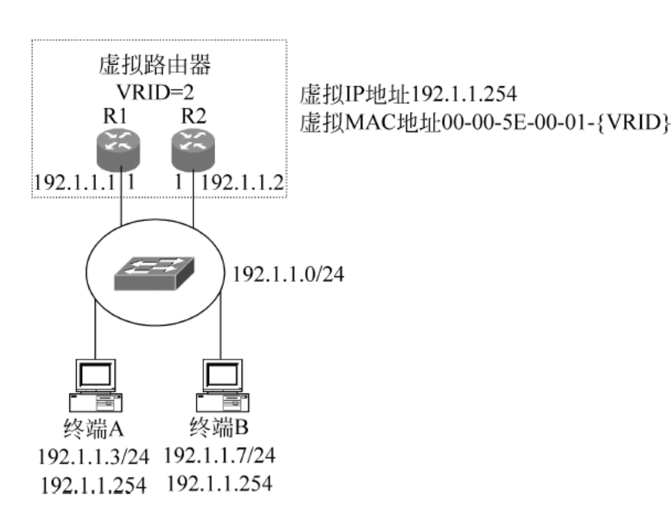


图 5.27 VRRP 工作环境

工作原理。每一个 VRRP 路由器连接以太网的接口可以分配多个 IP 地址，从这些 IP 地址中选择一个作为接口的基本 IP 地址，接口发送的 VRRP 报文以接口的基本 IP 地址作为 IP 分组的源 IP 地址。可以对虚拟路由器配置多个 IP 地址，这些 IP 地址称为虚拟 IP 地址，虚拟 IP 地址可以与为 VRRP 路由器接口配置的 IP 地址相同，如果某个 VRRP 路由器为某个接口配置的 IP 地址与为该接口所属的虚拟路由器配置的虚拟 IP 地址相同，该路由器称为 IP 地址拥有者。每一个虚拟路由器分配唯一的 8 位二进制数的虚拟路由器标识符（Virtual Router IDentifier，VRID），属于同一个虚拟路由器的多个 VRRP 路由器有着相同的虚拟路由器标识符。虚拟路由器对外有着唯一的 MAC 地址 00-00-5E-00-01-{VRID}，对于 VRID 为 2 的虚拟路由器，虚拟 MAC 地址为 00-00-5E-00-01-02。终端配置的默认网关地址必须是虚拟 IP 地址，对虚拟 IP 地址进行地址解析得到的结果必须是虚拟 MAC 地址，以虚拟 MAC 地址为目的 MAC 地址的 MAC 帧一定能够到达主路由器，只有主路由器转发封装在以虚拟 MAC 地址为目的 MAC 地址的 MAC 帧中的 IP 分组。

VRRP 需要解决的问题主要有以下三项。

- 在属于同一个虚拟路由器的多个 VRRP 路由器中产生主路由器；
- 一旦接收到终端发送的请求解析虚拟 IP 地址的 ARP 请求报文，虚拟路由器将虚拟 MAC 地址作为与虚拟 IP 地址绑定的 MAC 地址回送给终端；
- 以太网（严格地讲是所有支持广播的局域网）一定能够将以虚拟 MAC 地址为目的的 MAC 地址的 MAC 帧送达主路由器。

2. 路由器初始配置

对于图 5.27 所示的 VRRP 工作环境，路由器 R1 和路由器 R2 需要完成以下基本配置。

- 分别在路由器 R1 和路由器 R2 创建 VRID 为 2 的虚拟路由器，分别将路由器 R1 和路由器 R2 的接口 1 配置给 VRID 为 2 的虚拟路由器，使得路由器 R1 和路由器 R2 成为 VRID 为 2 的虚拟路由器的 VRRP 路由器；
- 分别为路由器 R1 和路由器 R2 的接口 1 分配 IP 地址 192.1.1.1/24 和 192.1.1.2/24，这两个接口的 IP 地址必须与它们连接的以太网的网络地址 192.1.1.0/24 一

致，由于路由器 R1 和路由器 R2 的接口 1 只分配了一个 IP 地址，该 IP 地址称为接口的基本 IP 地址；

- 为路由器 R1 和路由器 R2 的接口 1 分配优先级，优先级的范围为 1～254，主路由器用优先级 0 表示愿意主动放弃主路由器地位，IP 地址拥有者的优先级为 255。优先级值高的 VRRP 路由器在竞争主路由器时具有较高优先级；
- 为 VRID 为 2 的虚拟路由器分配虚拟 IP 地址 192.1.1.254。该 IP 地址成为连接在网络 192.1.1.0/24 上的终端的默认网关地址；
- 虚拟路由器根据 VRID＝2 生成虚拟 MAC 地址 00-00-5E-00-01-02。

3. VRRP 报文格式

VRRP 报文封装成 IP 分组的格式如图 5.28 所示，不直接将 VRRP 报文封装成 MAC 帧格式的主要原因是 VRRP 作用的网络可以是支持广播的任意网络，不一定是以太网。IP 分组的源 IP 地址是发送 VRRP 报文的接口的基本 IP 地址，对于路由器 R1 接口 1 发送的 VRRP 报文，其源 IP 地址为 192.1.1.1，目的 IP 地址是组播地址 224.0.0.18。所有 VRRP 路由器将以该组播地址为目的地址的 IP 分组提交给 VRRP 实体。VRRP 报文对应的协议字段值是 112。VRRP 报文中给出发送该 VRRP 报文的接口所属的虚拟路由器的 VRID、该接口的优先级、分配给虚拟路由器的虚拟 IP 地址等。VRRP 只有一种类型报文——通告报文。

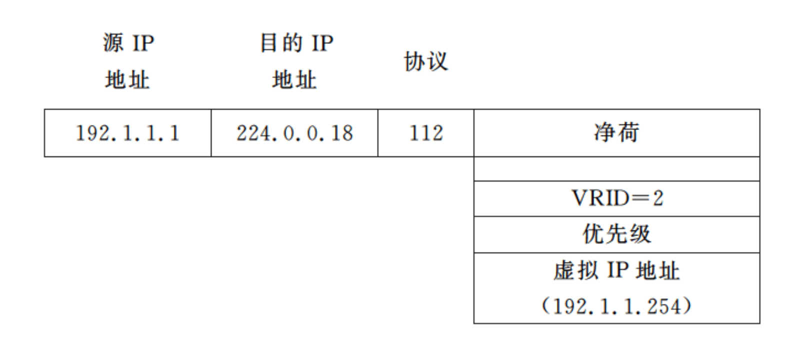


图 5.28　VRRP 报文格式

如果 VRRP 作用的网络是以太网，图 5.28 所示的 IP 分组将封装成 MAC 帧，该 MAC 帧的源 MAC 地址是发送接口所属虚拟路由器对应的虚拟 MAC 地址，对于路由器 R1 接口 1，源 MAC 地址是 00-00-5E-00-01-02，目的 MAC 地址是组播地址 224.0.0.18 对应的 MAC 组地址。根据组播地址 224.0.0.18 求出对应的 MAC 组地址的过程如图 5.29 所示。

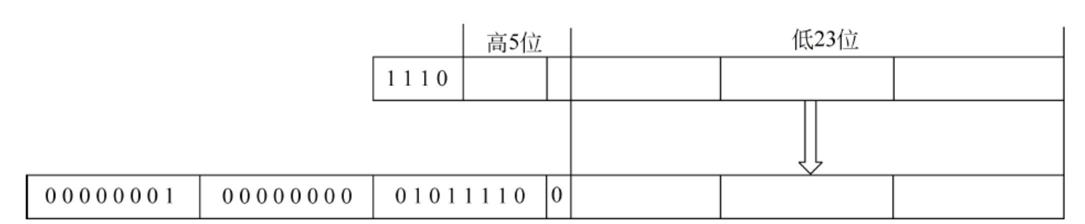


图 5.29　IP 组播地址映射到 MAC 组地址过程

从图 5.29 中可以看出，映射后的 MAC 地址的高 25 位固定为 00000001、00000000、01011110 和 0，低 23 位等于组播地址的低 23 位。因此，组播地址 224.0.0.18 对应的 MAC 组地址为 01-00-5E-00-00-12。由于组播地址中用于标识组播组的地址有 28 位，因此，标识组播组的组播地址中的高 5 位在映射过程中没有使用，这就使得组播地址和 MAC 组地址之间的映射不是唯一的，32 个不同的组播地址有可能映射为同一个 MAC 组地址。

4. 主路由器产生过程

路由器状态转换过程如图 5.30 所示，每一个 VRRP 路由器启动后，处于初始化状态，如果该 VRRP 路由器是 IP 地址拥有者，该 VRRP 路由器立即成为主路由器，并立即发送图 5.28 所示的 VRRP 报文，然后，周期性地发送 VRRP 报文。如果某个 VRRP 路由器不是 IP 地址拥有者，该 VRRP 路由器立即成为备份路由器，启动 Master_Down_Timer，等待接收主路由器发送的 VRRP 报文。

任何路由器接收到 VRRP 报文后，依序进行下列检查。

- 判别接收该 VRRP 报文的接口是否属于 VRRP 报文中 VRID 指定的虚拟路由器；
- 根据 VRRP 报文中的 VRID 确定虚拟路由器，判别路由器为该虚拟路由器配置的虚拟 IP 地址是否与 VRRP 报文中给出的虚拟 IP 地址相同。

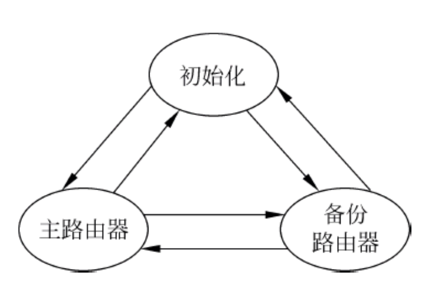


图 5.30 路由器状态转换过程

上述检查中只要有一项不匹配，路由器将丢弃该 VRRP 报文。

如果主路由器接收到 VRRP 报文，而且 VRRP 报文中的优先级大于主路由器为接收该 VRRP 报文的接口配置的优先级，或者虽然 VRRP 报文中的优先级等于主路由器为接收该 VRRP 报文的接口配置的优先级，但 VRRP 报文的源 IP 地址大于主路由器接收该 VRRP 报文的接口的基本 IP 地址，该主路由器立即转换为备份路由器，停止发送 VRRP 报文，启动 Master_Down_Timer，等待新的主路由器发送 VRRP 报文。

备份路由器接收到主路由器发送的 VRRP 报文后，根据备份路由器的工作方式对 VRRP 报文进行处理，如果备份路由器配置为允许抢占方式，且发现 VRRP 报文中的优先级小于备份路由器为接收该 VRRP 报文的接口配置的优先级，备份路由器立即转换为主路由器，并立即发送 VRRP 报文，然后，周期性地发送 VRRP 报文。如果备份路由器配置为不允许抢占方式，或者发现 VRRP 报文中的优先级大于或等于备份路由器为接收该 VRRP 报文的接口配置的优先级，刷新 Master_Down_Timer。

如果某个备份路由器的 Master_Down_Timer 溢出，表示主路由器已经失效，该备份路由器立即转换为主路由器，并立即发送 VRRP 报文，然后，周期性地发送 VRRP 报文。有可能因为网络拥塞导致主路由器发送的 VRRP 报文不能及时到达备份路由器，因而使备份路由器误认为主路由器失效而重新开始主路由器选择过程，为了避免发生这种情况，Master_Down_Timer 溢出时间大于 3×主路由器 VRRP 报文发送间隔。

5. 主路由器和备份路由器功能

1）主路由器功能

- 必须对请求解析虚拟 IP 地址的 ARP 请求报文做出响应；
- 必须对封装在以虚拟 MAC 地址为目的 MAC 地址的 MAC 帧中的 IP 分组进行转发操作；
- 在成为主路由器时，立即发送将所有虚拟 IP 地址绑定到虚拟 MAC 地址的 ARP 报文，使得网络内的所有终端将默认网关地址与虚拟 MAC 地址绑定在一起。

2）备份路由器功能

- 不对请求解析虚拟 IP 地址的 ARP 请求报文做出响应；
- 丢弃接收到的以虚拟 MAC 地址为目的地址的 MAC 帧；
- 丢弃接收到的以虚拟 IP 地址为目的地址的 IP 分组。

6. 虚拟 IP 地址解析过程

如果终端在 ARP 缓存中找不到与默认网关地址绑定的 MAC 地址，会发送一个请求解析该默认网关地址的 ARP 请求报文，该 ARP 请求报文在终端所连接的网络中广播，连接在该网络上的所有 VRRP 路由器都接收到该 ARP 请求报文，但只有主路由器对该 ARP 请求报文做出响应，并在 ARP 响应报文中将虚拟 MAC 地址与默认网关地址绑定在一起。终端发送给默认网关的 IP 分组封装在以终端 MAC 地址为源 MAC 地址，虚拟 MAC 地址为目的 MAC 地址的 MAC 帧中，只有主路由器对封装在这样 MAC 帧中的 IP 分组进行转发操作，其他 VRRP 路由器即使接收到该 MAC 帧，也将丢弃该 MAC 帧。

7. 交换机转发表更新过程

如果将图 5.27 中以太网扩展为图 5.31 所示以太网结构，在路由器 R2 成为主路由器后，以太网中各个交换机的转发表需要生成表 5.5 所示的转发项，否则，可能导致终端发送给默认网关的 MAC 帧在以太网中广播的情况。为了在各个交换机中生成表 5.5 所示的转发项，当路由器 R2 成为主路由器时，立即发送一个 VRRP 报文，该 VRRP 报文最终被封装成以虚拟 MAC 地址 00-00-5E-00-01-02 为源 MAC 地址，以组地址 01-00-5E-00-00-12 为目的 MAC 地址的 MAC 帧，该 MAC 帧在以太网中广播，如图 5.31 所示，以太网中所有交换机都接收到该 MAC 帧，通过地址学习，在转发表中建立表 5.5 所示的转发项。路由器 R2 通过定期发送 VRRP 报文定期刷新各个交换机中虚拟 MAC 地址对应的转发项，使得各个交换机一直在转发表中维持该转发项。

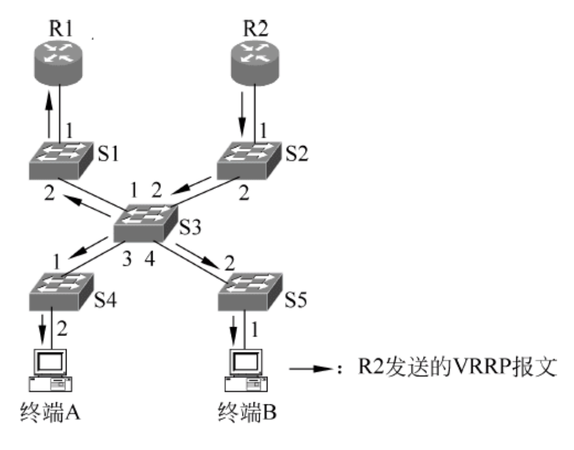


图 5.31 以太网结构

表 5.5 交换机转发表

MAC 地址	转发端口
交换机 S1	
00-00-5E-00-01-02	端口 2
交换机 S2	
00-00-5E-00-01-02	端口 1
交换机 S3	
00-00-5E-00-01-02	端口 2
交换机 S4	
00-00-5E-00-01-02	端口 1
交换机 S5	
00-00-5E-00-01-02	端口 2

8. 负载均衡

图 5.27 所示的 VRRP 工作环境能够解决容错问题,但无法实现负载均衡,为了实现负载均衡,采用图 5.32 所示的 VRRP 工作环境。创建两个 VRID 分别为 2 和 3 的虚拟路由器,同时将路由器 R1 和路由器 R2 连接以太网的接口分配给两个虚拟路由器,为 VRID 为 2 的虚拟路由器分配虚拟 IP 地址 192.1.1.1,使得路由器 R1 因为是 IP 地址拥有者而自然成为 VRID 为 2 的虚拟路由器中的主路由器。为 VRID 为 3 的虚拟路由器分配虚拟 IP 地址 192.1.1.2,使得路由器 R2 因为是 IP 地址拥有者而自然成为 VRID 为 3 的虚拟路由器中的主路由器。将一半连接在网络 192.1.1.0/24 上的终端(图 5.32 中的终端 A)的默认网关地址配置成 VRID 为 2 的虚拟路由器对应的虚拟 IP 地址 192.1.1.1,将另一半连接在网络 192.1.1.0/24 上的终端(图 5.32 中的终端 B)的默认网关地址配置成 VRID 为 3 的虚拟路由器对应的虚拟 IP 地址 192.1.1.2,这样,连接在网络 192.1.1.0/24 上的终端,一半将路由器 R1 作为默认网关,另一半将路由器 R2 作为默认网关,一旦某个路由器发生故障,另一个路由器将自动作为所有终端的默认网关,既实现了容错,又实现了负载均衡。

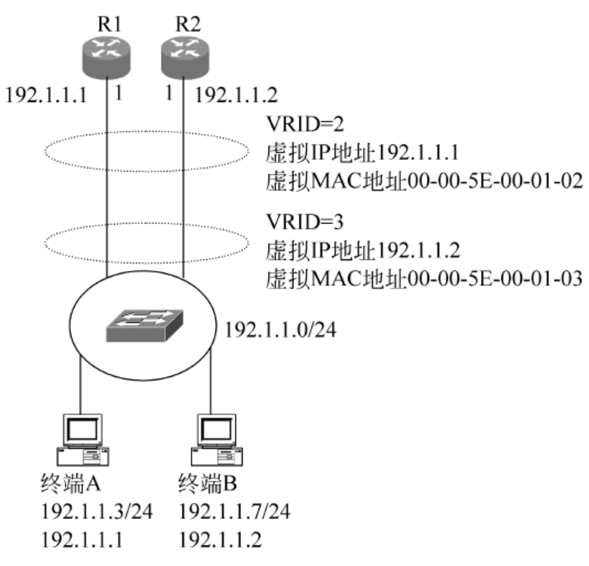


图 5.32 均衡负载的 VRRP 工作环境

5.4.3　VRRP 应用实例

1. 网络结构与基本配置

图 5.33 是图 5.26 的简化版，为了实现容错和负载均衡，对网络进行如下配置。

- 根据图 5.33 所示配置信息分别为路由器 R1 和路由器 R2 的两个接口配置 IP 地址和子网掩码，完成路由器接口 IP 地址和子网掩码配置后，路由器 R1 和路由器 R2 自动生成图 5.33 所示的路由表，路由表中给出用于指明通往路由器直接连接的网络的传输路径的路由项；
- 创建 VRID 分别为 2 和 3 的两个虚拟路由器，并将路由器 R1 接口 1 和路由器 R2 接口 1 分配给 VRID 为 2 的虚拟路由器，并将路由器 R1 接口 2 和路由器 R2 接口 2 分配给 VRID 为 3 的虚拟路由器，VRID 为 2 的虚拟路由器对应的虚拟 MAC 地址为 00-00-5E-00-01-02，VRID 为 3 的虚拟路由器对应的虚拟 MAC 地址为 00-00-5E-00-01-03；
- 为 VRID 为 2 的虚拟路由器分配虚拟 IP 地址 192.1.1.254，这使得路由器 R1 成为 VRID 为 2 的虚拟路由器的主路由器，为 VRID 为 3 的虚拟路由器分配虚拟 IP 地址 192.1.2.253，这使得路由器 R2 成为 VRID 为 3 的虚拟路由器的主路由器；
- 连接在网络 192.1.1.0/24 上的终端配置默认网关地址 192.1.1.254，连接在网络 192.1.2.0/24 上的终端配置默认网关地址 192.1.2.253。

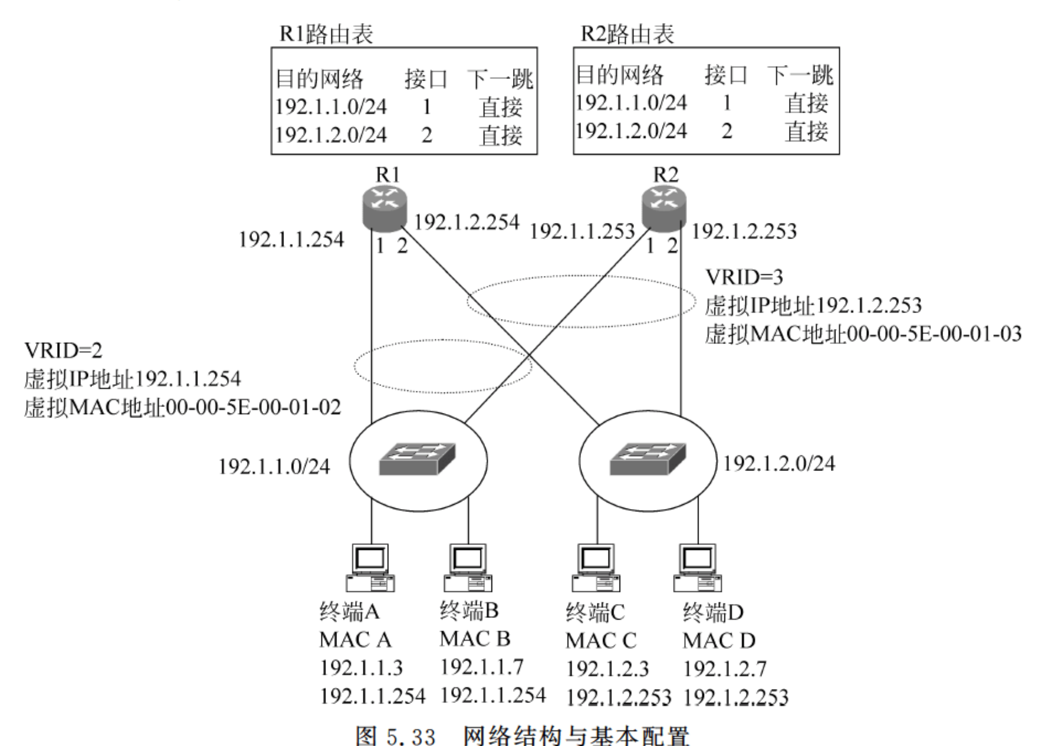


图 5.33　网络结构与基本配置

2. 生成主路由器和转发项

路由器 R1 因为是虚拟 IP 地址 192.1.1.254 的 IP 地址拥有者，自然成为 VRID 为 2 的

虚拟路由器的主路由器，在成为主路由器时，一是通过发送 VRRP 报文，在网络 192.1.1.0/24 各个交换机中建立将目的 MAC 地址为 00-00-5E-00-01-02 的 MAC 帧转发给路由器 R1 接口 1 的转发项。同时，通过在网络 192.1.1.0/24 中广播将虚拟 IP 地址 192.1.1.254 与虚拟 MAC 地址 00-00-5E-00-01-02 绑定的 ARP 报文，在连接在网络 192.1.1.0/24 上的所有终端的 ARP 缓存中建立 IP 地址 192.1.1.254 与 MAC 地址 00-00-5E-00-01-02 的绑定。同样的，在路由器 R2 成为 VRID 为 3 的虚拟路由器的主路由器后，在网络 192.1.2.0/24 各个交换机中建立将目的 MAC 地址为 00-00-5E-00-01-03 的 MAC 帧转发给路由器 R2 接口 2 的转发项。在连接在网络 192.1.2.0/24 上的所有终端的 ARP 缓存中建立 IP 地址 192.1.2.253 与 MAC 地址 00-00-5E-00-01-03 的绑定。

3. IP 分组传输过程

如果终端 A 需要向终端 D 发送 IP 分组，首先获取终端 D 的 IP 地址 192.1.2.7，构建源 IP 地址为 192.1.1.3、目的 IP 地址为 192.1.2.7 的 IP 分组。通过判别终端 A 和终端 D 所在网络的网络地址(192.1.1.0/24 和 192.1.2.0/24)发现终端 A 和终端 D 不在同一个网络，终端 A 需要将 IP 分组发送给默认网关。终端 A 从 ARP 缓存中获取默认网关地址 192.1.1.254 绑定的 MAC 地址 00-00-5E-00-01-02，构建以终端 A 的 MAC 地址 MAC A 为源 MAC 地址，以 MAC 地址 00-00-5E-00-01-02 为目的 MAC 地址的 MAC 帧，网络地址 192.1.1.0/24 保证将该 MAC 帧转发给路由器 R1，路由器 R1 由于是 VRID 为 2 的虚拟路由器的主路由器，必须对封装在以虚拟 MAC 地址 00-00-5E-00-01-02 为目的 MAC 地址的 MAC 帧中的 IP 分组进行转发操作。路由器 R1 从 MAC 帧中分离出 IP 分组，用 IP 分组的目的 IP 地址 192.1.2.7 匹配路由器 R1 路由表中的路由项，发现和路由项＜192.1.2.0/24，2，直接＞匹配，下一跳为直接表明目的终端连接在接口 2 连接的网络上，通过 ARP 地址解析过程获取与目的 IP 地址 192.1.2.7 绑定的 MAC 地址 MAC D，构建以接口 2 所属的虚拟路由器对应的虚拟 MAC 地址 00-00-5E-00-01-03 为源 MAC 地址，以终端 D 的 MAC 地址 MAC D 为目的 MAC 地址的 MAC 帧，通过网络地址 192.1.2.0/24 将该 MAC 帧转发给终端 D，终端 D 从 MAC 帧中分离出 IP 分组，完成终端 A 至终端 D IP 分组的传输过程。

当终端 D 向终端 A 发送 IP 分组时，终端 D 先将 IP 分组转发给默认网关——路由器 R2，实现了路由器 R1 和路由器 R2 的负载均衡。当其中一个路由器发生故障，另一个路由器将作为连接在两个网络上的终端的默认网关。

习题

5.1 为什么说 IP 是一种网际协议？IP 实现连接在不同传输网络上的终端之间通信的技术基础是什么？

5.2 为什么为每一个路由器接口分配 IP 地址？

5.3 作为中继系统，转发器、网桥和路由器有何区别？

5.4 解释不能用网桥实现两个分别连接在以太网和 ATM 网络的终端之间通信的原因。

5.5　解释路由器和网桥的主要区别。

5.6　何为默认网关？终端配置默认网关的作用是什么？

5.7　路由器实现不同类型的传输网络互连的技术基础是什么？

5.8　路由器主要由几部分组成？如何实现IP分组的转发过程？

5.9　IP地址分为几类？各类如何表示？它们的主要特点是什么？

5.10　简述IP地址和MAC地址之间的不同，及各自的作用。

5.11　为什么需要无分类编址？它对路由项聚合和子网划分带来什么好处？

5.12　什么是最长前缀匹配算法？在什么条件下需要使用最长前缀匹配算法？

5.13　子网掩码255.255.255.0代表什么意思？如果某一网络的子网掩码为255.255.255.248，该网络能够连接多少主机？

5.14　以下地址中的哪一个地址和网络前缀86.32/12匹配，说明理由。

A. 86.33.224.123　　B. 86.79.65.216

C. 86.58.119.74　　D. 86.68.206.154

5.15　以下网络前缀中的哪一个和地址2.52.90.140匹配，说明理由。

A. 0/4　　B. 32/4　　C. 4/6　　D. 80/4

5.16　请辨认以下IP地址的类型。

(1) 128.36.199.3；

(2) 21.12.240.17；

(3) 183.194.76.253；

(4) 192.12.69.248；

(5) 89.3.0.1；

(6) 200.3.6.2。

5.17　一个3200b的TCP报文传到IP层，加上160b的首部后成为IP分组，下面的互连网络由两个局域网通过路由器连接起来，但第二个局域网的MTU=1200b，因此，IP分组必须在路由器进行分片。试问第二个局域网实际需要为上层传输多少比特的数据？

5.18　假定传输层将包含20B首部和2048B数据的TCP报文递交给IP层，源终端至目的终端传输路径需要经过两个网络，其中第一个网络的MTU=1024B，第二个网络的MTU=512B，IP首部是20B，给出到达目的终端时分片后的IP分组序列，并计算出每一片的净荷字节数和片偏移。

5.19　路径MTU是端到端传输路径所经过网络中最小的MTU，假定源终端能够发现路径MTU，并以路径MTU作为源终端封装IP分组的依据，根据5.18题的参数，给出到达目的终端时分片后的IP分组序列，并计算出每一片的净荷字节数和片偏移。

5.20　有人说“ARP向网络层提供了转换地址的服务，应该属于数据链路层”，为什么说这种说法是错误的？

5.21　ARP缓冲器中每一项的寿命是10～15min，简述寿命太长或者太短可能出现的问题。

5.22　如果重新设计IP地址时，将IP地址设计为48位，能否通过IP地址和MAC地址之间的一一对应关系消除ARP地址解析过程？

5.23　设某路由器建立了如下路由表(这三列分别是目的网络、子网掩码和下一跳路由

器，若直接交付，则最后一列给出输出接口）。

128.96.39.0	255.255.255.128	接口 0
128.96.39.128	255.255.255.128	接口 1
128.96.40.0	255.255.255.128	R2
192.4.153.0	255.255.255.192	R3
默认		R4

现收到 5 个 IP 分组，其目的 IP 地址如下：

（1）128.96.39.10；

（2）128.96.40.12；

（3）128.96.40.151；

（4）192.4.153.17；

（5）192.4.153.90。

试分别计算出下一跳路由器或输出接口。

5.24　某单位分配到一个 B 类 IP 地址，其网络地址为 124.250.0.0，该单位有 4000 多台机器，分布在 16 个不同的地点，如果选用的子网掩码为 255.255.255.0，试分别给每个地点分配一个网络地址，并根据网络地址计算出每个地点可分配的 IP 地址范围。

5.25　一个 IP 分组的数据长度为 4000B（固定长度首部），需要经过一个 MTU 为 1500B 的网络，试问应当划分为几个数据片？每一个数据片的数据字段长度、片偏移字段和 MF 标志为何值？

5.26　IP 分组中的首部检验和只检验 IP 分组首部，这样做的好处是什么？坏处是什么？IP 分组首部检错码为什么不采用 CRC？

5.27　一个自治系统有 5 个局域网，其连接如图 5.34 所示，LAN 2～LAN 5 上的主机数分别为 91、150、3 和 15，该自治系统分配到的 IP 地址块为 30.138.118.0/23，试给出每一个局域网的地址块（包括前缀）。

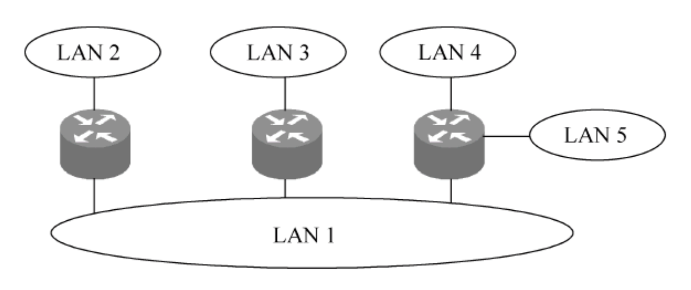


图 5.34　题 5.27 图

5.28　对如下 4 个地址块进行最大可能的聚合。

212.56.132.0/24

212.56.133.0/24

212.56.134.0/24

212.56.135.0/24

5.29　根据图 5.35 所示的网络地址配置，给出路由器 R1、路由器 R2 和路由器 R3 的路由表。如果要求路由器 R2 中的路由项最少，如何调整网络地址配置？并根据调整后的

网络地址配置，给出路由器 R1、路由器 R2 和路由器 R3 的路由表。

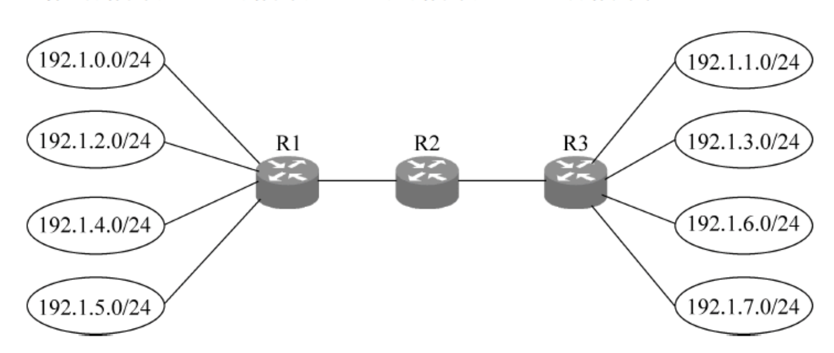


图 5.35　题 5.29 图

5.30　根据图 5.36 所示的互连网络结构，为每一个局域网分配合适的网络前缀地址（假定 CIDR 地址块为 192.77.33.0/24，图中每一个局域网旁边标明的数字是该局域网的主机数）。

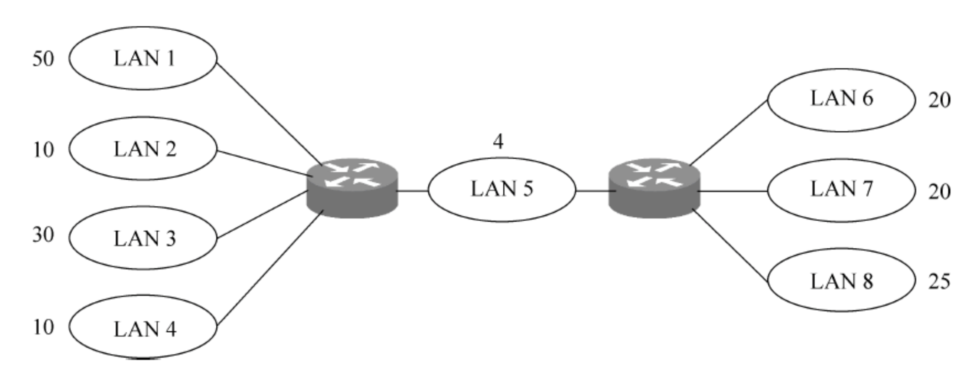


图 5.36　题 5.30 图

5.31　某单位分配到一个地址块 136.23.12.64/26，现在需要进一步划分为 4 个一样大的子网，试问：

（1）每个子网的网络前缀有多长？

（2）每个子网有多少地址？

（3）每个子网的地址块是什么？

（4）每个子网可分配给主机的最小和最大地址是什么？

5.32　网络结构如图 5.37 所示，给出的 CIDR 地址块是 192.1.1.64/26，确定每一个子网的网络地址，将最大可用地址分配给路由器连接对应子网的接口，给出路由器 R1、R2 的路由表。

5.33　互连网络结构如图 5.38 所示。

① 补齐图中终端和路由器的配置信息，包括路由表。使其能够实现终端 A 和终端 B 之间的 IP 分组传输。

② 以①补齐的配置信息为基础，给出终端 A 至终端 B IP 分组传输过程中涉及的所有 MAC 帧，并给出这些 MAC 帧的源和目的 MAC 地址（假定终端和路由器的 ARP 缓冲器为空）。

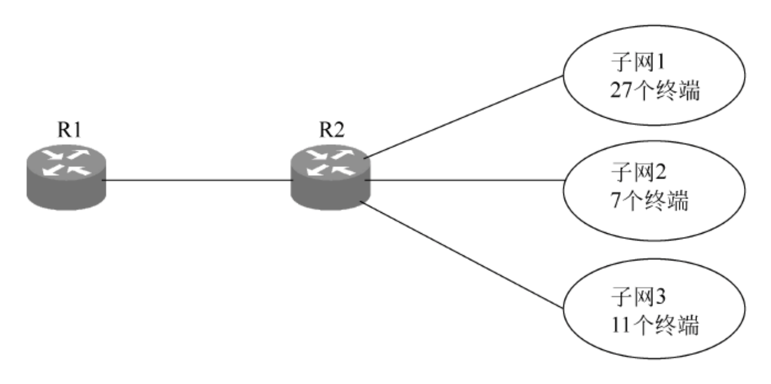


图 5.37　题 5.32 图

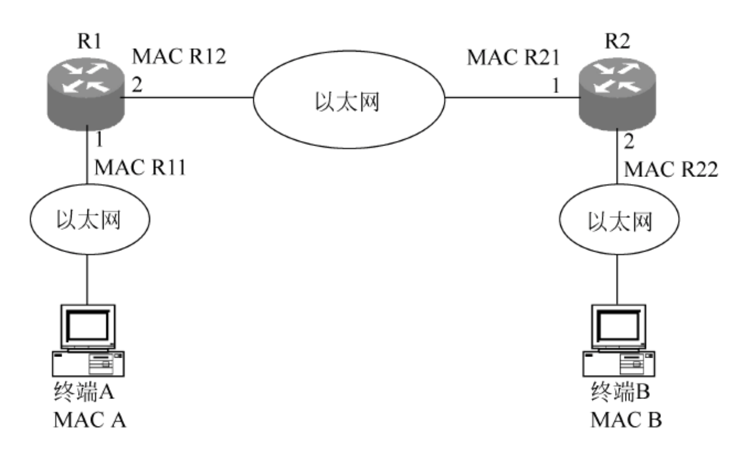


图 5.38　题 5.33 图

5.34　VRRP 的作用是什么？

5.35　简述主路由器转换为备份路由器的条件。

5.36　简述备份路由器转换为主路由器的条件。

5.37　对于图 5.33 所示网络结构，如果要求连接在网络 192.1.1.0/24 和 192.1.2.0/24 上终端，各有一半以路由器 R1、路由器 R2 为默认网关，给出实现这一功能所需的 VRRP 配置。

第6章 路由协议

IP 分组通过逐跳转发实现端到端传输过程，逐跳转发的关键是路由表，每一跳路由器通过检索路由表得知通往目的终端传输路径上的下一跳路由器，可以通过手工配置静态路由项的方式在每一个路由器中创建路由表，但这种方式存在很大缺陷，因此，需要一种能够根据网络拓扑结构在每一个路由器上自动创建路由表的协议，这种协议就是路由协议。

6.1 路由项分类

6.1.1 直连路由项

图 6.1 中每一个路由器连接三个网络，为路由器接口配置的 IP 地址和子网掩码确定了该接口所连接的网络的网络地址。如果为路由器 R1 接口 1 配置 IP 地址和子网掩码 192.1.1.254/24，通过对 192.1.1.254 和 255.255.255.0 进行“与”操作，得到结果 192.1.1.0，因此得出路由器 R1 接口 1 所连接的网络的网络地址是 192.1.1.0/24。以此得出路由器 R1 接口 2 所连接的网络的网络地址是 192.1.4.0/30（将 192.1.4.1 和 255.255.255.252进行“与”操作结果），接口 3 所连接的网络的网络地址是 192.1.5.0/30

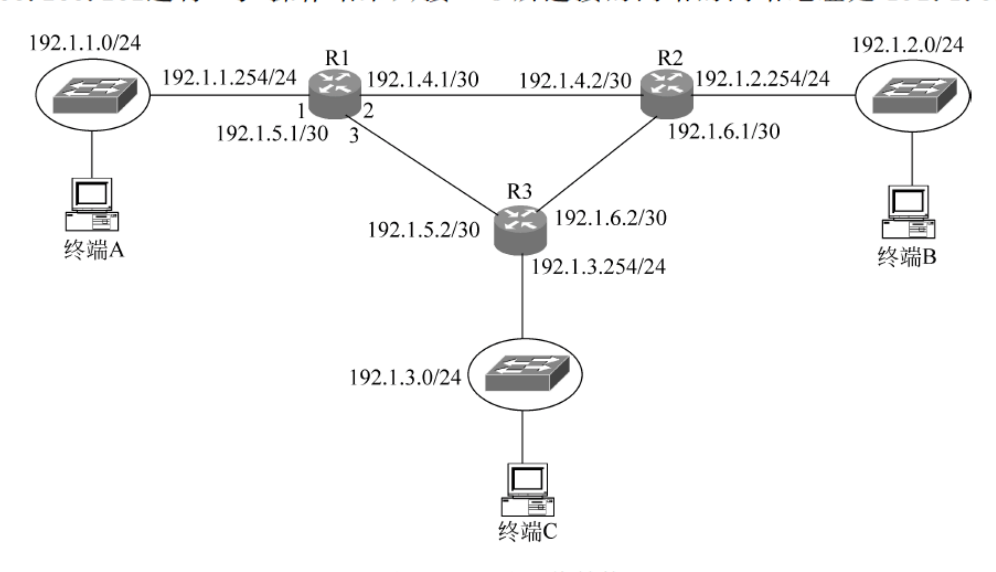


图 6.1 互连网络结构

(将 192.1.5.1 和 255.255.255.252 进行“与”操作结果)。路由器同样给出用于指明通往这些直接连接的网络的传输路径的路由项,这些路由项称为直连路由项。在为路由器接口配置 IP 地址和子网掩码后,路由器自动根据接口配置的 IP 地址和子网掩码生成直连路由项。对于路由器 R1,在完成接口的 IP 地址和子网掩码配置后,生成如表 6.1 所示的直连路由项。网络地址 192.1.4.0/30 只包含 4 个 IP 地址(192.1.4.0～192.1.4.3),其中 192.1.4.0 是网络地址(主机字段为全 0),192.1.4.3 是直接广播地址(主机字段为全 1)。因此,只包含两个有效 IP 地址 192.1.4.1 和 192.1.4.2,这种类型的网络地址是无分类编址中有效 IP 地址数量最少的网络地址,一般用于用点对点链路互连两个路由器的连接方式中。

表 6.1 路由器 R1 直连路由项

目的网络	输出接口	下一跳
192.1.1.0/24	1	直接
192.1.4.0/30	2	直接
192.1.5.0/30	3	直接

如果两个网络直接连接在同一个路由器上,只要完成路由器接口的 IP 地址和子网掩码配置,路由器将自动生成用于指明通往这两个网络的传输路径的路由项,对于连接在这两个网络上的终端,只要完成 IP 地址、子网掩码和默认网关地址配置,就可实现相互通信。对于图 6.2 所示互连网络结构,在完成路由器接口 IP 地址和子网掩码配置,终端 A 和终端 B 的 IP 地址、子网掩码和默认网关地址配置后,终端 A 和终端 B 之间就可进行 IP 分组双向传输过程。

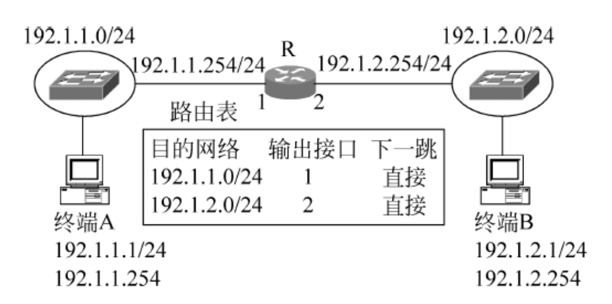


图 6.2 单个路由器的互连网络结构

6.1.2 静态路由项

对于图 6.2 所示的多个网络直接与同一个路由器连接的互连网络结构,在完成路由器接口的 IP 地址和子网掩码配置后,路由器路由表中能够自动生成用于指明通往这些直接连接的网络的传输路径的直连路由项,路由器根据直连路由项能够完成连接在不同网络上的终端之间的 IP 分组转发操作。但对于图 6.1 所示的网络结构,如果需要实现 IP 分组终端 A 至终端 B 的传输过程,由于终端 B 连接的网络没有和路由器 R1 直接连接,因此,路由器 R1 的路由表中没有用于指明通往网络 192.1.2.0/24 的传输路径的直连路由项,路由器 R1 由于无法确定通往网络 192.1.2.0/24 传输路径上的下一跳,将丢弃所有目的 IP 地址属于网络地址 192.1.2.0/24 的 IP 分组。因此,对于所有没有和某个路由器直接连接的网络,该路由器必须生成用于指明通往这些网络的传输路径的路由项,否则,该路由器将丢弃以连接

在这些网络上的终端为目的终端的 IP 分组。对于图 6.1 中的路由器 R1，没有和其直接连接的网络有三个，分别是网络 192.1.2.0/24、192.1.3.0/24 和 192.1.6.0/30。如果路由器 R1 需要转发以属于这些网络地址的 IP 地址为目的 IP 地址的 IP 分组，路由器 R1 必须在路由表中生成用于指明通往这三个网络的传输路径的路由项。如果采用手工配置静态路由项的方式，首先需要确定路由器 R1 至这三个网络的最短路径，然后求出路由器 R1 至这三个网络的最短路径上的下一跳路由器，及下一跳路由器连接路由器 R1 的接口的 IP 地址，根据这些信息得出路由器 R1 用于指明通往这三个网络的传输路径的路由项。对于网络 192.1.2.0/24，路由器 R1 通往该网络的最短路径是 R1→R2→192.1.2.0/24（传输路径经过的路由器跳数最少），且路由器 R2 连接路由器 R1 的接口的 IP 地址是 192.1.4.2，得出用于指明通往网络 192.1.2.0/24 的传输路径的路由项为＜192.1.2.0/24，2，192.1.4.2＞，其中，192.1.2.0/24 是目的网络的网络地址，2 是输出接口编号，192.1.4.2 是下一跳地址。同样得出用于指明通往网络 192.1.3.0/24 的传输路径的路由项为＜192.1.3.0/24，3，192.1.5.2＞。由于路由器 R1 存在两条经过的路由器跳数相同的通往网络 192.1.6.0/30 的传输路径，可以在两条传输路径中任选一条，这里，选择传输路径 R1→R2→192.1.6.0/30 作为通往网络 192.1.6.0/30 的传输路径，并因此生成路由项＜192.1.6.0/30，2，192.1.4.2＞。由此可以得出表 6.2 所示的路由器 R1 用于指明通往所有 6 个网络的传输路径的路由项，类型 C 表示的是路由器自动生成的直连路由项，类型 S 表示的是手工配置的静态路由项。需要强调的是，如果下一跳 IP 地址是 192.1.4.2，则输出接口肯定是路由器 R1 连接网络192.1.4.0/30 的接口。同样，如果下一跳 IP 地址是 192.1.5.2，输出接口肯定是路由器 R1 连接网络 192.1.5.0/30 的接口。

表 6.2　路由器 R1 完整路由表

路由项类型	目的网络	输出接口	下一跳
C	192.1.1.0/24	1	直接
C	192.1.4.0/30	2	直接
C	192.1.5.0/30	3	直接
S	192.1.2.0/24	2	192.1.4.2
S	192.1.3.0/24	3	192.1.5.2
S	192.1.6.0/30	2	192.1.4.2

6.1.3　动态路由项

在确定互连网络拓扑结构和完成路由器接口 IP 地址和子网掩码配置的前提下，通过在路由器运行路由协议生成的与互连网络拓扑结构一致的、用于指明通往互连网络中所有网络的传输路径的路由项称为动态路由项，使用动态路由项这一术语的主要目的是为了区别手工配置的静态路由项。

1. 路由协议

每一个路由器通过和其他路由器相互交换路由消息，发现与互连网络拓扑结构一致的、通往互连网络中所有网络的最短路径，并据此生成用于指明通往互连网络中所有网络的最

短路径的路由项。路由协议就是一组用于规范路由消息的格式、路由器之间路由消息交换过程、路由器对路由消息的处理流程的规则。目前，存在多种路由协议，虽然所有路由协议的作用都是为互连网络中的每一个路由器找出通往互连网络中所有网络的最短路径，但不同路由协议对最短路径的定义、对路由消息格式和内容的约定等都是不同的。

2. 路径距离

所谓最短路径，就是路径距离最小的传输路径，如果某个路由器存在多条通往某个网络的传输路径，如图 6.1 中，路由器 R1 存在两条通往网络 192.1.2.0/24 的传输路径，一是 R1→R2→192.1.2.0/24，另一条是 R1→R3→R2→192.1.2.0/24，在这些传输路径中选择路径距离最小的传输路径。路径距离可以是传输路径经过的路由器跳数，也可以是其他衡量传输路径的参数，如传输路径的物理距离、传输路径经过的物理链路的带宽等。如果以传输路径经过的路由器跳数作为传输路径距离，传输路径 R1→R2→192.1.2.0/24 的距离为 1(Cisco 计算跳数时不包含传输路径起始路由器，以后路由协议计算传输路径跳数时与此习惯一致)，传输路径 R1→R3→R2→192.1.2.0/24 的距离为 2。距离最小的传输路径为最短路径。如果以传输路径经过的物理链路带宽作为传输路径距离，则首先需要定义将带宽换算成代价的计算公式(如计算公式：代价＝10^8/带宽，当带宽是 100Mb/s 时，得出代价是 1，当带宽是 10Mb/s 时，得出代价为 10)。计算传输路径距离时，需要累计传输路径经过的物理链路的代价和，如果互连 R1 和 R3、R3 和 R2 的物理链路的带宽为 100Mb/s，互连 R1 和 R2 的物理链路的带宽为 10Mb/s，路由器 R2 连接网络 192.1.2.0/24 的物理链路的带宽为 100Mb/s，则传输路径 R1→R2→192.1.2.0/24 的距离为 11，传输路径 R1→R3→R2→192.1.2.0/24 的距离为 3。由于路由协议要求代价必须是整数，因此，当物理链路带宽大于 100Mb/s 时，不能使用计算公式：代价＝108/带宽，而是需要为物理链路定义一个能够反映物理链路带宽的距离值。

6.1.4 静态路由项缺陷

手工配置静态路由项的缺陷是显然的，一是大型互连网络很难做到各个路由器配置的静态路由项一致，这主要因为很难保证用户为每一个路由器选择的通往特定网络的传输路径是一致的；二是互连网络的拓扑结构是动态变化的，通过手工改变各个路由器中的静态路由项来适应不断变化的互连网络拓扑结构是不现实的；三是为了容错，各个网络之间存在多条传输路径，理想的工作状态是多条传输路径同时承担网络之间的流量，以实现负载均衡，并在其中一条传输路径发生故障的情况下，迅速将由该传输路径承担的流量分摊到其他传输路径上，通过手工配置静态路由项来实现这一功能几乎是不可能的；四是大型互连网络中的路由器不仅很多，而且分布的物理区域很广泛，在每一个路由器上配置静态路由项的工作量是无法想象的。

6.2 路由协议基础

路由协议作用前，互连网络中的每一个路由器需要通过为接口配置 IP 地址和子网掩码自动生成直连路由项，路由协议在每一个路由器生成的直连路由项的基础上，为每一个路由

器动态生成用于指明通往所有没有和该路由器直接连接的网络的传输路径的路由项，由路由协议生成的路由项称为动态路由项。

6.2.1 路由协议分类

可以根据路由协议发现、计算最短路径的方式，也可以根据路由协议的作用范围对路由协议进行分类。

1. 距离向量路由协议和链路状态路由协议

根据路由协议交换路由消息和计算最短路径的方式可以将路由协议分为距离向量路由协议和链路状态路由协议。

1）距离向量路由协议

路由协议的功能是在每一个路由器自动生成的直连路由项的基础上，通过路由器之间交换路由消息，使得每一个路由器能够生成用于指明通往互连网络中没有和该路由器直接连接的网络的传输路径的路由项。距离向量路由协议要求每一个路由器定期向其相邻路由器公告全部路由项，由于每一项路由项用于指明通往某个网络或网络前缀相同的一组网络的传输路径，路由器拥有某项路由项，意味着该路由器已经建立通往目的网络字段指定的一个或一组网络的传输路径，当某个路由器接收到相邻路由器公告的路由消息，且该路由器没有路由消息中包含的某项路由项，意味着该路由器可以通过相邻路由器到达该路由项目的网络字段指定的一个或一组网络，该路由器就可创建一项路由项，目的网络字段值与路由消息中该路由项的目的网络字段值相同，下一跳为相邻路由器连接该路由器的接口的 IP 地址。通过相邻路由器之间不断交换路由消息，最终使互连网络中的所有路由器建立用于指明通往互连网络中所有没有和该路由器直接连接的网络的传输路径的路由项。之所以将该路由协议称为距离向量路由协议，是因为发送给相邻路由器的路由消息由一组路由项组成，且路由项格式为＜目的网络，距离＞。如果某个路由器可以通过多个相邻路由器到达某个网络，通过距离值在多条通往某个网络的传输路径中选择距离最短的传输路径。两个路由器相邻，意味着两个路由器存在连接在同一个网络上的接口，这样的接口也称为两个路由器的互连接口。目前常见的距离向量路由协议有路由信息协议（Routing Information Protocol，RIP）和边界网关协议（Border Gateway Protocol，BGP）。

2）链路状态路由协议

互连网络中的每一个网络必须和某个路由器直接连接，路由器的每一个接口连接一个网络。该网络或是末端网络，除了该路由器接口，不再连接其他路由器接口；或是互连路由器网络，两个或两个以上路由器存在连接该网络的接口，一组存在连接在同一个网络上的接口的路由器（称为相邻路由器）。每一个路由器可以通过一组链路状态来表示接口所连接的网络，链路状态为＜Router ID，Neighbor，Cost＞。Router ID 是某个路由器标识符，通常用该路由器其中一个接口的 IP 地址作为该路由器标识符。如果某个接口连接的是末端网络，Neighbor 是该网络的网络地址，如果某个接口连接的是互连路由器网络，Neighbor 是相邻路由器连接互连路由器网络的接口的 IP 地址。Cost 是根据路由器连接该网络的接口的带宽计算出的代价。每一个路由器通过一组链路状态表示和该路由器直接连接的网络的信息。当互连网络中的某个路由器获得所有其他路由器的链路状态信息，就可构建互连网络

的拓扑结构，并在此基础上计算出该路由器到达所有网络的最短路径。目前常见的链路状态路由协议有开放最短路径优先(Open Shortest Path First，OSPF)。

2. 内部网关协议和外部网关协议

一个大型互连网络中，无数个网络和互连网络的路由器被划分成了多个自治系统(Autonomous System，AS)，每一个自治系统通常由单一管理部门负责管理，运行相同的路由协议。但自治系统不是孤岛，必须由设备将自治系统互连在一起，这种用于互连自治系统的设备称为自治系统边界路由器(Autonomous System Boundary Router，ASBR)，这样一来，两个不属于同一自治系统的终端之间的传输过程涉及两个层次的传输路径，一是连接源终端所在网络的路由器如何找到一条通往连接源终端所在自治系统的自治系统边界路由器的传输路径，二是连接源终端所在自治系统的自治系统边界路由器如何找到一条通往连接目的终端所在自治系统的自治系统边界路由器的传输路径。前一条传输路径是由属于同一自治系统的路由器构成，后一条传输路径是由互连不同自治系统的自治系统边界路由器构成，如图 6.3 所示。把用于建立第一条传输路径的路由协议称为内部网关协议(Interior Gateway Protocol，IGP)，而将用于建立第二条传输路径的路由协议称为外部网关协议(External Gateway Protocol，EGP)。常用的内部网关协议有 RIP、OSPF，外部网关协议有 BGP。

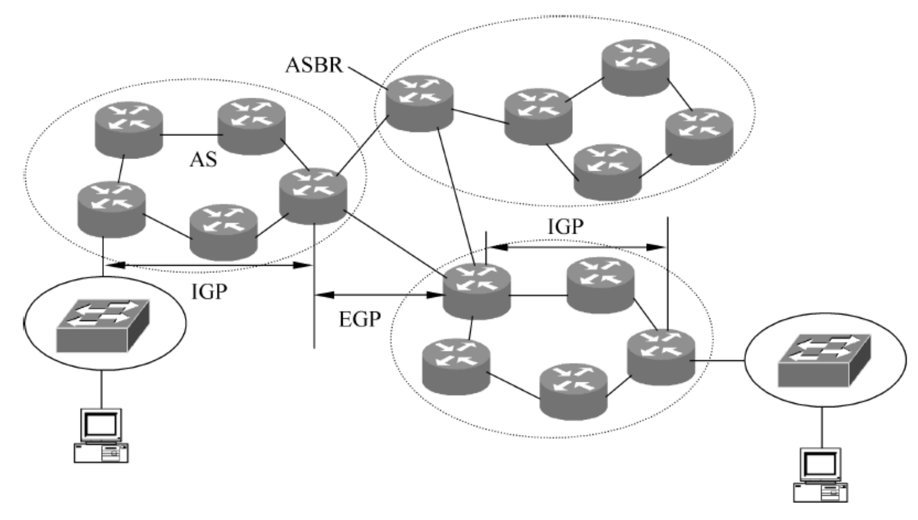


图 6.3 二层传输路径

6.2.2 路由协议要求

1. 建立完整路由

路由协议必须在每一个路由器中创建用于指明通往所有没有与其直接连接的网络的传输路径的路由项。

2. 选择最佳路由

必须在所有路由器中统一代价的含义，所有路由器选择的通往特定网络的传输路径是一致的。

3. 简单、开销小

路由协议创建路由项过程必须简单，路由器计算开销和网络传输路由消息开销必须要小，运行路由协议不会对路由器正常转发 IP 分组和网络传输 IP 分组产生较大影响。

4. 实时反映互连网络拓扑结构的变化

一旦互连网络拓扑结构发生变化，如增加新的网络、某个路由器发生故障、某条物理链路发生故障等，每一个路由器中的路由项必须及时更新，以适应变化后的互连网络拓扑结构。

5. 具有稳定性

稳定性体现在两个方面，一是在互连网络拓扑结构没有发生变化的情况下，各个路由器中的路由项保持稳定；二是对于任何互连网络拓扑结构，每一个路由器创建的路由项是固定的、可预测的。

6. 快速收敛

收敛是指在互连网络拓扑结构不变的前提下，每一个路由器都创建了用于指明通往所有没有和其直接连接的网络的传输路径的路由项，而且互连网络中所有路由器创建的路由项是一致的。快速收敛一是指在互连网络拓扑结构发生变化时，所有路由器能够快速创建适应变化后的互连网络拓扑结构，且相互一致的路由项；二是当某个路由器重新启动时，能够快速创建和其他路由器一致的、用于指明通往所有没有和其直接连接的网络的传输路径的路由项。

6.2.3 距离向量路由协议

1. 距离向量路由协议创建路由表过程

1）建立直连路由项

为图 6.1 中路由器 R1、路由器 R2 和路由器 R3 的每一个接口配置 IP 地址和子网掩码后，路由器 R1、路由器 R2 和路由器 R3 路由表中自动生成表 6.3～表 6.5 所示的直连路由项，直连路由项的距离为 0，表示直连路由项经过的路由器跳数为 0。

表 6.3　路由器 R1 直连路由项

类型	目的网络	输出接口	距离	下一跳
C	192.1.1.0/24	1	0	直接
C	192.1.4.0/30	2	0	直接
C	192.1.5.0/30	3	0	直接

表 6.4　路由器 R2 直连路由项

类型	目的网络	输出接口	距离	下一跳
C	192.1.2.0/24	1	0	直接
C	192.1.6.0/30	2	0	直接
C	192.1.4.0/30	3	0	直接

表 6.5 路由器 R3 直连路由项

类型	目的网络	输出接口	距离	下一跳
C	192.1.3.0/24	1	0	直接
C	192.1.5.0/30	2	0	直接
C	192.1.6.0/30	3	0	直接

2) 定期交换路由消息

路由器 R1 分别与路由器 R2 和路由器 R3 相邻，因此，定期相互交换路由消息，路由器 R2 发送给路由器 R1 的路由消息如下：{(192.1.2.0/24,0)(192.1.6.0/30,0)(192.1.4.0/30,0)192.1.4.2}，其中包含路由器 R2 的全部直连路由项和路由器 R2 连接路由器 R1 的接口的 IP 地址，直连路由项(192.1.2.0/24,0)中 192.1.2.0/24 表示目的网络，0 表示路由器 R2 通往目的网络 192.1.2.0/24 的传输路径的距离为 0。由于路由器 R1 没有与网络 192.1.2.0/24和网络 192.1.6.0/30 直接连接，且通过路由器 R2 发送的路由消息获知，路由器 R2 能够到达这些网络，因此，路由器 R1 发现经过路由器 R2 到达这些网络的传输路径，并在路由表中创建用于指明通往网络 192.1.2.0/24 和 192.1.6.0/30 的传输路径的路由项，对于路由器 R1，通往网络 192.1.2.0/24 和 192.1.6.0/30 的传输路径上的下一跳是路由器 R2，下一跳 IP 地址应该是路由消息给出的路由器 R2 连接路由器 R1 的接口的 IP 地址 192.1.4.2。路由器 R1 增加用于指明通往网络 192.1.2.0/24 和 192.1.6.0/30 的传输路径的路由项后的路由表如表 6.6 所示，表中用 D 表示路由项类型是路由协议创建的动态路由项。距离 1 是路由器 R1 到达网络 192.1.2.0/24 的传输路径经过的路由器跳数，由于路由器 R2 到达网络 192.1.2.0/24 的传输路径的距离为 0，而路由器 R1 到达网络 192.1.2.0/24 的传输路径是路由器 R1 至路由器 R2 传输路径＋路由器 R2 到达网络 192.1.2.0/24的传输路径，需要在路由器 R2 到达网络 192.1.2.0/24 的传输路径的距离上加 1。同样，路由器 R3 向路由器 R1 发送路由消息{(192.1.3.0/24,0)(192.1.5.0/30,0)(192.1.6.0/30,0)192.1.5.2}，路由器 R1 根据路由器 R3 发送的路由消息生成用于指明通往网络 192.1.3.0/24 和 192.1.6.0/30 的传输路径的路由项，由于路由器 R1 的路由表中已经存在用于指明通往网络 192.1.6.0/30 的传输路径的路由项，根据最短路径原则，路由器 R1 选择距离最小的路由项作为最终路由项，在两项路由项距离相等的情况下，路由器 R1 任选一项路由项作为最终路由项，路由器 R1 生成的完整路由表如表 6.7 所示。同样路由器 R1 也向路由器 R2 和路由器 R3 发送路由消息，在建立表 6.7 所示的完整路由表后，路由器 R1 发送给路由器 R2 和路由器 R3 的路由消息分别如下：{(192.1.1.0/24,0)(192.1.4.0/30,0)(192.1.5.0/30,0)(192.1.2.0/24,1)(192.1.3.0/24,1)(192.1.6.0/30,1)192.1.4.1}和{(192.1.1.0/24,0)(192.1.4.0/30,0)(192.1.5.0/30,0)(192.1.2.0/24,1)(192.1.3.0/24,1)(192.1.6.0/30,1)192.1.5.1}，两个路由消息中不同的是用于作为下一跳路由器地址的 IP 地址。路由器 R2 和路由器 R3 也根据路由器 R1 发送的路由消息创建路由项。经过路由器之间多次相互交换路由消息，路由器 R1、路由器 R2 和路由器 R3 最终生成用于指明通往所有网络的传输路径的路由项。这个时候，各个路由器的路由表已经收敛。

表 6.6　路由器 R1 路由表

类型	目的网络	输出接口	距离	下一跳
C	192.1.1.0/24	1	0	直接
C	192.1.4.0/30	2	0	直接
C	192.1.5.0/30	3	0	直接
D	192.1.2.0/24	2	1	192.1.4.2
D	192.1.6.0/30	2	1	192.1.4.2

表 6.7　路由器 R1 完整路由表

类型	目的网络	输出接口	距离	下一跳
C	192.1.1.0/24	1	0	直接
C	192.1.4.0/30	2	0	直接
C	192.1.5.0/30	3	0	直接
D	192.1.2.0/24	2	1	192.1.4.2
D	192.1.6.0/30	2	1	192.1.4.2
D	192.1.3.0/24	3	1	192.1.5.2

2. 距离向量路由协议特性

1）周期性广播全部路由项

每一个路由器必须向其相邻路由器定期发送路由消息，由于无法确定某个路由器接口连接的网络中存在哪些相邻路由器，因此，路由器在某个接口连接的网络上广播路由消息。路由消息中给出该路由器的全部路由项，发送路由消息的间隔时间决定收敛时间和路由消息传输开销，减小发送路由消息的间隔时间，会减小收敛时间，但会增加路由消息的传输开销，容易导致网络发生拥塞。加大发送路由消息的间隔时间，会增加收敛时间，但会减少路由消息的传输开销。

2）容易发生路由环路

由于每一个路由器根据相邻路由器发送的路由消息来生成路由项，有可能导致路由环路。路由环路是指一条成环的传输路径，如图 6.1 中，路由器 R1 通往某个特定网络的传输路径的下一跳是路由器 R2，路由器 R2 通往该网络的传输路径的下一跳是路由器 R3，路由器 R3 通往该网络的传输路径的下一跳是路由器 R1。这样各个路由器通往该网络的传输路径构成环路。

3）实时性差

当网络拓扑结构发生变化时，重新收敛各个路由器的路由表的时间较长。

4）设置触发机制

除了周期性发送路由消息，必须在发现有路由项发生改变的情况下，立即向其相邻路由器发送路由消息，以此加快相邻路由器路由表的收敛速度。

5）设置无效定时器

如果某项路由项根据相邻路由器发送的路由消息创建，当该相邻路由器发生故障时，该路由项应该无效，无效定时器用于确定没有接收到该相邻路由器发送的路由消息的最长时

间间隔，如果在无效定时器规定的时间间隔内，一直没有接收到该相邻路由器发送的路由消息，可以断定该相邻路由器已经发生了故障，该时间间隔一般是3×相邻路由器路由消息发送周期。

6.2.4 链路状态路由协议

1. 链路状态路由协议创建路由表过程

1）建立各个路由器的链路状态

为图6.1中路由器R1、路由器R2和路由器R3的每一个接口配置IP地址和子网掩码后，路由器R1、路由器R2和路由器R3之间通过交换Hello报文获得每一个接口所连接的链路的状态。表6.8是每一个路由器建立的链路状态。Cost字段是根据路由器接口带宽换算出的代价，换算公式为Cost=10^8/接口传输速率，这里假定路由器R1连接路由器R2的链路的传输速率为10Mb/s，其他链路的传输速率为100Mb/s。

表6.8 路由器链路状态

Router ID	Neighbor	Cost
路由器R1链路状态		
R1	192.1.1.0/24	1
R1	192.1.4.2(R2)	10
R1	192.1.5.2(R3)	1
路由器R2链路状态		
R2	192.1.2.0/24	1
R2	192.1.4.1(R1)	10
R2	192.1.6.2(R3)	1
路由器R3链路状态		
R3	192.1.3.0/24	1
R3	192.1.5.1(R1)	1
R3	192.1.6.1(R2)	1

2）泛洪链路状态

每一个路由器将自身链路状态封装成链路状态通告后，以泛洪方式传输给互连网络中的所有其他路由器。泛洪方式传输过程如下：始发路由器通过所有接口广播链路状态通告，某个路由器接收到链路状态通告后，首先向广播该链路状态通告的路由器发送确认应答，然后判别是否已经接收过该链路状态通告，如果是第一次接收该链路状态通告，从除接收该链路状态通告接口以外的所有其他接口广播该链路状态通告。如果已经接收过该链路状态通告，则丢弃该链路状态通告。链路状态通告中包含始发路由器标识符和序号，路由器每发送一个新的链路状态通告，序号递增，序号最大的链路状态通告是始发路由器发送的最新的链路状态通告。某个路由器接收到某个始发路由器发送的链路状态通告后，判别该链路状态通告中携带的序号是否大于该路由器为始发路由器保留的序号，如果条件成立，将链路状态通告携带的序号作为为始发路由器保留的序号，表明该路由器第一次接收到该链路状态通告。如果条件不成立，表明路由器已经接收过该链路状态通告。图6.4给出路由器

R1 泛洪链路状态通告过程，当路由器 R3 接收到路由器 R2 转发的链路状态通告时，由于路由器 R2 已经接收过路由器 R1 发送的链路状态通告，且这两个链路状态通告的始发路由器和序号都相同，所以路由器 R3 丢弃路由器 R2 转发的链路状态通告。

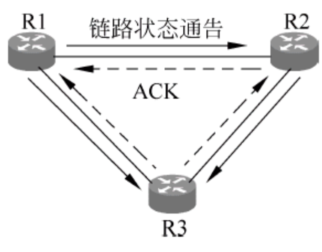


图 6.4 路由器 R1 泛洪链路状态通告过程

3）构建链路状态数据库

当所有路由器泛洪自身链路状态后，互连网络中的每一个路由器建立表 6.8 所示的链路状态库，该链路状态库描述了互连网络的拓扑结构。

4）计算最短路径树

下面以构建路由器 R1 为树根的最短路径树为例，讨论根据链路状态库构建最短路径树的算法。令 $D(v)$ 为源结点（路由器 R1）到达结点 v 的距离，它是从源结点沿着某一路径到达结点 v 所经过的链路的代价之和，$L(i,j)$ 为结点 i 至结点 j 的距离。

① 以 R1 为树根，求出各个结点和根结点之间距离。

$$D(v)=\begin{cases} L(R1,v) & \text{若结点 } v \text{ 与 R1 直接相连} \\ \infty & \text{若结点 } v \text{ 与 R1 不直接相连} \end{cases}$$

② 找出与根结点距离最短的结点（假定为结点 w），将该结点连接到以 R1 为根的树上，并重新对剩下的结点计算到达根结点的距离，$D(v)=\text{MIN}\{D(v),D(w)+L(w,v)\}$。

③ 重复步骤②，直到所有结点都连接到以源结点为根的树上。

表 6.9 给出了构建路由器 R1 为根的最短路径树的每一步。将根路由器 R1 连接到最短路径树上，路由器 R1 到达自身的距离为 0。找出与路由器 R1 直接连接的结点和网络放入备份结点和网络中，根据表 6.8 所示的链路状态库，与路由器 R1 直接连接的结点和网络有路由器 R2、路由器 R3 和 192.1.1.0/24。距离分别是 10、1 和 1。选择距离最小的结点或网络直接连接到根结点上。

选择了将结点路由器 R3 直接连接到根结点后，重新计算各个结点和网络到达根结点的距离，计算出路由器 R2 经过路由器 R3 到达根结点的距离为 2。$D(2)=D(3)+L(3,2)1+1=2$。由于路由器 R2 直接到达路由器 R1 的距离大于路由器 R2 经过路由器 R3 到达路由器 R1 的距离，结点路由器 R2 必须连接到最短路径树路由器 R3 分枝上。经过表 6.9 所示的步骤，最终生成图 6.5 所示的以路由器 R1 为根的最短路径树。根据图 6.5 所示的最短路径树，得出表 6.10 所示的路由器 R1 路由表。

表 6.9 以路由器 R1 为根的最短路径树生成过程

最短路径树	备份结点和网络	说　明
(R1,R1,0)	(R1,192.1.1.0/24,1) (R1,R2,10) (R1,R3,1)	从备份结点和网络中选择到达路由器 R1 距离最短的结点或网络连接到根结点上，第一步选择网络 192.1.1.0/24
(R1,R1,0) (R1,192.1.1.0/24,1)	(R1,R2,10) (R1,R3,1)	选择路由器 R3 连接到根结点上

续表

最短路径树	备份结点和网络	说　明
(R1,R1,0) (R1,192.1.1.0/24,1) (R1,R3,1)	(R1,R2,2)(根据(R3,R2,1)计算出路由器 R2 到达路由器 R1 的距离为 2) (R1,192.1.3.0/24,2)(根据(R3,192.1.3.0/24,1)计算出网络 192.1.3.0/24到达路由器 R1 的距离为 2)	根据路由器 R3 重新计算各个结点和网络到达路由器 R1 的距离。选择网络 192.1.3.0/24 连接到最短路径树的路由器 R3 分枝上
(R1,R1,0) (R1,192.1.1.0/24,1) (R1,R3,1) (R3,192.1.3.0/24,1)	(R1,R2,2)	选择路由器 R2 连接到最短路径树的路由器 R3 分枝上
(R1,R1,0) (R1,192.1.1.0/24,1) (R1,R3,1) (R3,192.1.3.0/24,1) (R3,R2,1)	(R1,192.1.2.0/24,3)(根据(R3,R2,1)和(R2,192.1.2.0/24,1)计算出网络 192.1.2.0/24 到达路由器 R1 的距离为 3)	根据路由器 R2 重新计算各个结点和网络到达路由器 R1 的距离。将网络 192.1.2.0/24 连接到最短路径树路由器 R2 分枝上

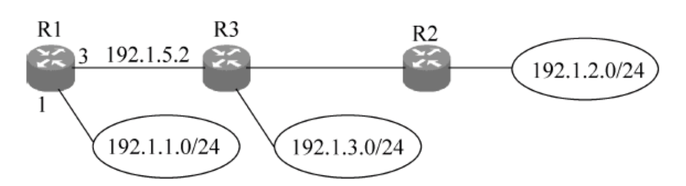


图 6.5　以路由器 R1 为根的最短路径树

表 6.10　路由器 R1 完整路由表

类型	目的网络	输出接口	距离	下一跳
C	192.1.1.0/24	1	0	直接
C	192.1.4.0/30	2	0	直接
C	192.1.5.0/30	3	0	直接
D	192.1.2.0/24	3	3	192.1.5.2
D	192.1.3.0/24	3	2	192.1.5.2

2. 链路状态路由协议特性

1）快速收敛

通过互连网络中各个路由器泛洪链路状态通告,互连网络中的每一个路由器很快建立链路状态库,并根据链路状态库构建以自己为根的最短路径树。

2）消除路由环路

由于每一个路由器有着相同的链路状态库,并根据链路状态库构建以自己为根的最短路径树,各个路由器根据以自己为根的最短路径树生成的路由表是不会产生路由环路的。

3）实时性好

一旦某个路由器的链路状态发生变化,该路由器通过泛洪链路状态通告及时向互连网

络中的所有其他路由器通报这种变化，使得其他路由器能够及时更新链路状态库，并重新构建以自己为根的最短路径树。

4）实现负载均衡

由于每一个路由器都具有描述互连网络拓扑结构的链路状态库，可以计算出到达某个特定网络的所有传输路径，并根据流量分配策略将传输给该特定网络的流量分配到多条不同的传输路径上。

5）传输开销较大

由于每一个路由器都需要将自己的链路状态封装成链路状态通告，并以泛洪方式将链路状态通告传输给互连网络中的所有其他路由器，因此这种传输链路状态通告的方式给网络增加较多流量。

6）计算复杂度高

根据链路状态库构建以自己为根的最短路径树的算法是一种计算复杂度很高的算法，因此，每一个路由器根据链路状态库构建以自己为根的最短路径树的过程会占用路由器大量的计算能力，会对路由器转发 IP 分组的能力造成影响。

3. 例题解析

【例 6.1】 互连网络结构及互连路由器的链路类型如图 6.6 所示，链路类型与代价的关系如表 6.11 所示，假定 NET1 和 NET2 都是快速以太网，求出终端 A 至终端 B 的最短路径，路由器 R1 和路由器 R5 对应 NET1 和 NET2 的动态路由项。

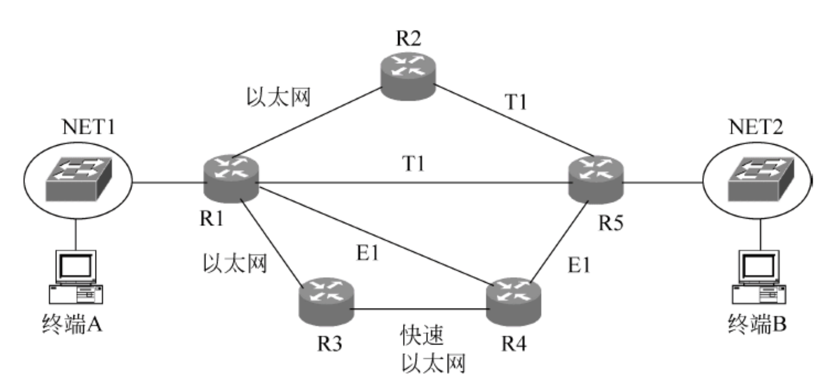


图 6.6 互连网络结构和链路类型

表 6.11 链路类型和代价

链路类型	传输速率	代 价
快速以太网	100Mb/s	$(10^8)/(100\times10^6)=1$
以太网	10Mb/s	$(10^8)/(10\times10^6)=10$
E1	2.048Mb/s	$(10^8)/(2.048\times10^6)=48$
T1	1.544Mb/s	$(10^8)/(1.544\times10^6)=64$

【解析】 求终端 A 和终端 B 之间的最短路径实际上是求路由器 R1 和路由器 R5 之间的最短路径，路由器 R1 和路由器 R5 之间路径和距离如下：

R1→R3→R4→R5 ，距离＝10＋1＋48＝59。

R1→R4→R5,距离=48+48=96。

R1→R5,距离=64。

R1→R2→R5,距离=10+64=74。

显然终端A至终端B的最短路径=终端A→R1→R3→R4→R5→终端B。

路由器到达某个网络的距离是路由器通往该网络的传输路径经过的所有路由器输出链路的代价之和,因此,路由器R1到达网络NET2的距离是路由器R1到达路由器R5的距离+路由器R5连接NET2的链路的代价=59+1=60。直连路由项的距离通常假定为0。路由器R1和路由器R5的路由表分别如表6.12和表6.13所示。

表6.12 路由器R1路由表

目的网络	下一跳	距离
NET1	直接	0
NET2	R3	60

表6.13 路由器R5路由表

目的网络	下一跳	距离
NET1	R4	60
NET2	直接	0

6.3 RIP

路由信息协议(Routing Information Protocol,RIP)是一种基于距离向量的路由协议,在路由器通过配置接口的IP地址和子网掩码而自动生成的直连路由项的基础上,通过相邻路由器之间不断交换路由消息,最终在所有路由器中建立通往所有网络的最短路径。

6.3.1 RIP消息格式

RIP消息格式如图6.7所示,主要给出发送RIP消息的路由器的路由项,每一项路由项中的IP地址和子网掩码给出路由项的目的网络,距离给出该路由器到达目的网络所经过的路由器跳数,如果接收该RIP消息的路由器采用某项路由项,对于该路由器,发送RIP消息的路由器将成为该项路由项中的下一跳路由器,发送RIP消息的接口的IP地址成为下一跳IP地址。但在一些特殊情况下,对于该项路由项,可能存在比发送RIP消息的路由器更好的下一跳,这种情况下,RIP消息通过下一跳地址指定该路由器。因此,大多数情况下,RIP消息中每一项路由项中有用的信息是IP地址、子网掩码和距离。

RIP消息被封装成UDP报文,该UDP报文通过源端口和目的端口号520指明净荷是RIP消息。封装RIP消息的IP分组的源IP地址是发送该RIP消息的接口的IP地址,一旦接收该RIP消息的路由器采用了RIP消息包含的某项路由项,该路由器将用封装RIP消息的IP分组的源IP地址作为该项路由项的下一跳IP地址。

图6.7所示的RIP消息格式是RIP响应消息格式,用于周期性公告全部路由项。当某个路

由器刚启动时，也可向相邻路由器发送 RIP 请求消息，要求相邻路由器立即发送其路由表包含的全部路由项，接收到 RIP 请求消息的路由器将立即发送 RIP 响应消息。RIP 消息通过命令字段标识两种不同的消息类型。以后讨论时，如果不指定 RIP 消息类型，表示是 RIP 响应消息。

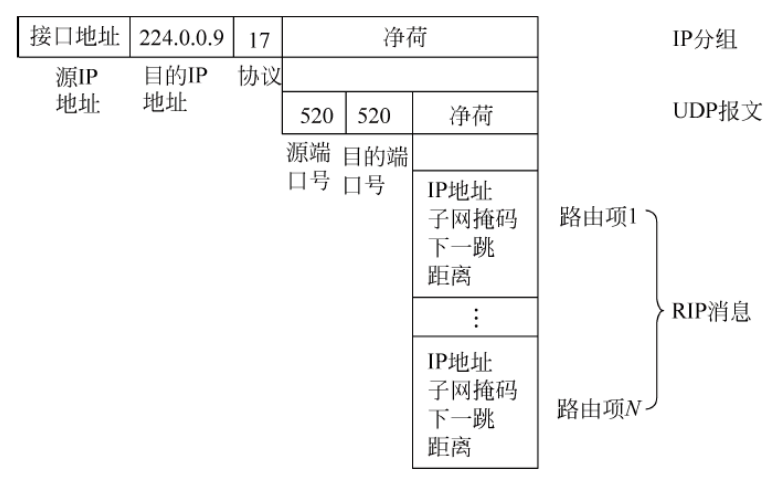


图 6.7　RIP 消息格式及封装过程

6.3.2　RIP 工作过程

1. 基本思路

RIP 的工作思路如下：用 D(i,j)表示路由器 i 到达网络 j 的距离，如果某个路由器 i 直接连接某个网络 j，则该路由器到达该网络的距离最短，距离为 0，D(i,j)＝0。如果某个路由器 i 没有直接和某个网络 j 连接，则必须找到一个 D(k,j)为最短路径距离的相邻路由器 k，使得 D(i,j)＝ D(k,j)＋1，且 D(i,j)为路由器 i 到达网络 j 的最短路径的距离，即如果路由器 i 的相邻路由器集合＝$\{k_1,k_2,\cdots,k_N\}$，则 D(i,j)＝MIN[D(i,k_i)＋D(k_i,j)]，$k_i \in \{k_1,k_2,\cdots,k_N\}$。RIP 用 16 表示无穷大距离，用于指示不可达的传输路径距离，如果 D(i,j)＝16，表明路由器 i 和网络 j 之间不存在传输路径。由此可以得出适用 RIP 的是端到端传输路径的最大跳数小于等于 15 的互连网络，因此，RIP 只适用于较小规模的自治系统。

2. 定期交换路由消息

RIP 工作基础是路由器通过配置接口 IP 地址和子网掩码自动生成的直连路由项。初始时，路由器路由表中只包含直连路由项，通过相邻路由器之间不断交换路由消息，每一个路由器逐渐建立用于指明通往和其没有直接连接的网络的传输路径的路由项。

每一个路由器只和相邻路由器交换路由消息，两个路由器相邻指的是两个路由器存在连接在同一个网络的接口，因此，两个路由器可以直接经过该网络实现通信。由于存在多个路由器连接在同一个网络的情况，因此，从某个接口发送出去的路由消息，必须被所有有接口连接在该网络的路由器接收，因此，封装路由消息的 IP 分组的目的 IP 地址是表明这样一组路由器的组播地址：224.0.0.9，源 IP 地址是发送路由消息的接口的 IP 地址。

路由消息中给出该路由器已经建立的路由项，路由项格式为＜目的网络，距离＞(虽然路由项中包含下一跳地址，但大部分情况下以封装路由消息的 IP 分组的源 IP 地址作为下一跳地址)。路由器启动 RIP 进程时，每一个路由器的路由表中只包含直连路由项，因此，

一开始每一个路由器只能向其相邻路由器发送包含直连路由项的路由消息。

随着路由器之间不断交换路由消息，每一个路由器逐渐建立用于指明通往所有网络的最短路径的路由项，由于互连网络是不断变化的，因此，路由表中的路由项也是不断变化的，为了使所有路由器及时感知变化的互连网络，某个路由器一旦发现路由表中有路由项发生变化，立即向其相邻路由器公告这一变化。为了确定最短路径的工作状态，每一个路由器必须定期向其相邻路由器发送路由消息。

3. 路由器处理路由消息流程

当某个路由器 Y 接收到其相邻路由器 X 发送给它的路由消息时，根据路由消息中的路由项＜N,D(X,N)＞确定路由器 Y 到达网络 N 的最短路径的过程如下：

① D(Y,N)＝D(X,N)＋1；

② 如果路由器 Y 的路由表中没有用于指明通往网络 N 的最短路径的路由项，说明传输路径 Y→X 和 X→N 是路由器 Y 发现的第一条通往网络 N 的传输路径，以该传输路径为最短路径，生成对应的路由项，目的网络＝N，距离＝D(Y,N)，下一跳＝路由器 X(用封装路由消息的 IP 分组的源 IP 地址标识)，设置定时器。

③ 如果路由器 Y 的路由表中已经存在用于指明通往网络 N 的最短路径的路由项，且该路由项指明的通往网络 N 的最短路径和传输路径 Y→X 和 X→N 不同，根据最短路径原则，路由器 Y 将选择距离较短的传输路径作为最短路径，因此，如果路由器 Y 中存在路由项＜N,D′(Y,N),X′＞,X′≠X 且 D(Y,N)＜D′(Y,N)，路由器 Y 将传输路径 Y→X 和 X→N 作为最短路径，用新的路由项＜N,D (Y,N),X＞(目的网络＝N，距离＝D(Y,N)，下一跳＝路由器 X)取代原来的路由项，并重新设置定时器，否则保持原来的路由项不变。

④ 如果路由器 Y 的路由表中已经存在路由项＜N,D′(Y,N),X＞，说明路由器 Y 通往网络 N 的最短路径就是传输路径 Y→X 和 X→N，重新设置定时器，如果 D(Y,N)≠D′(Y,N)，说明 X→N 的最短路径距离已经发生变化，必须在路由项中用 D(Y,N)取代 D′(Y,N)，以反映当前 Y→N 最短路径的实际距离，如果 D(Y,N)≥16，则将路由项的距离设置为 16，表示该路由项指明的传输路径已经不可达。

⑤ 如果 D(X,N) ＝16，意味着 X→N 传输路径已不存在，如果路由器 Y 中路由项指明的通往网络 N 的最短路径包含传输路径 X→N，即目的网络＝N 的路由项中，下一跳路由器＝X，路由器 Y 将删除或停止使用该路由项(将该路由项距离设置成 16)。

路由表中每一项路由项都有定时器，重新设置定时器(也称刷新定时器)表示重新开始定时器计时，如果总是在定时器溢出前进行重新设置定时器操作，定时器将不会溢出，一旦定时器溢出，将该路由项的距离设置为 16，表明该路由项指定的最短路径已经不可达。

【例 6.2】 假定路由器 Y 的路由表如表 6.14 所示，接收到的来自相邻路由器 X 的路由消息如表 6.15 所示，求出路由器 Y 处理表 6.15 所示的路由消息中路由项后的路由表。

表 6.14 路由器 Y 路由表

目的网络	距离	下一跳路由器
N2	3	X
N3	6	A
N4	5	X
N5	7	X

表 6.15 路由器 X 发送的路由消息

目的网络	距离
N1	3
N2	6
N3	3
N4	4
N5	16

【解析】 对于路由消息中的第一项路由项，由于路由器 Y 路由表中没有目的网络＝N1 的路由项，在路由表中增添路由项＜N1，3＋1，X＞。

对于路由消息中的第二项路由项，由于路由器 Y 路由表中存在目的网络＝N2 且下一跳路由器＝X 的路由项，用新的距离 7 取代老的距离 3。

对于路由消息中的第三项路由项，虽然路由器 Y 路由表中存在目的网络＝N3，下一跳路由器＝A 的路由项，由于以路由器 X 为下一跳路由器的传输路径距离(4)小于以路由器 A 为下一跳路由器的传输路径距离，用较短距离的传输路径取代原来的传输路径。

对于路由消息中的第四项路由项，由于无论距离，还是下一跳路由器都和路由器 Y 中已经存在的路由项相同，对路由器 Y 中的路由项不做任何修改，只是重新设置定时器。

对于路由消息中的第五项路由项，由于其距离为 16，而且路由器 Y 中已经存在目的网络为 N5 且下一跳路由器为 X 的路由项，将该路由项的距离设置成 16，表明该路由项指定的传输路径不可达。

处理完表 6.15 所示的路由消息中路由项后的路由器 Y 路由表如表 6.16 所示。

表 6.16 路由器 Y 处理路由消息后的路由表

目的网络	距离	下一跳路由器
N1	4	X
N2	7	X
N3	4	X
N4	5	X
N5	16	X

6.3.3 RIP 建立路由表实例

下面以图 6.8 所示互连网络结构为例，讨论路由器 R5 通过 RIP 建立用于指明通往所有网络的最短路径的路由项的过程。

1. 路由器建立初始路由表

首先通过为图 6.8 所示互连网络中路由器的各个接口配置 IP 地址和子网掩码，使各个路由器自动生成只包含直连路由项的初始路由表，这里为了简单起见，初始路由表中的路由项及以后建立的路由项只和图 6.8 特地指定的 4 个网络有关，因此，只有路由器 R1、路由器 R3、路由器 R5、路由器 R7 建立如表 6.17～表 6.20 所示的初始路由表。

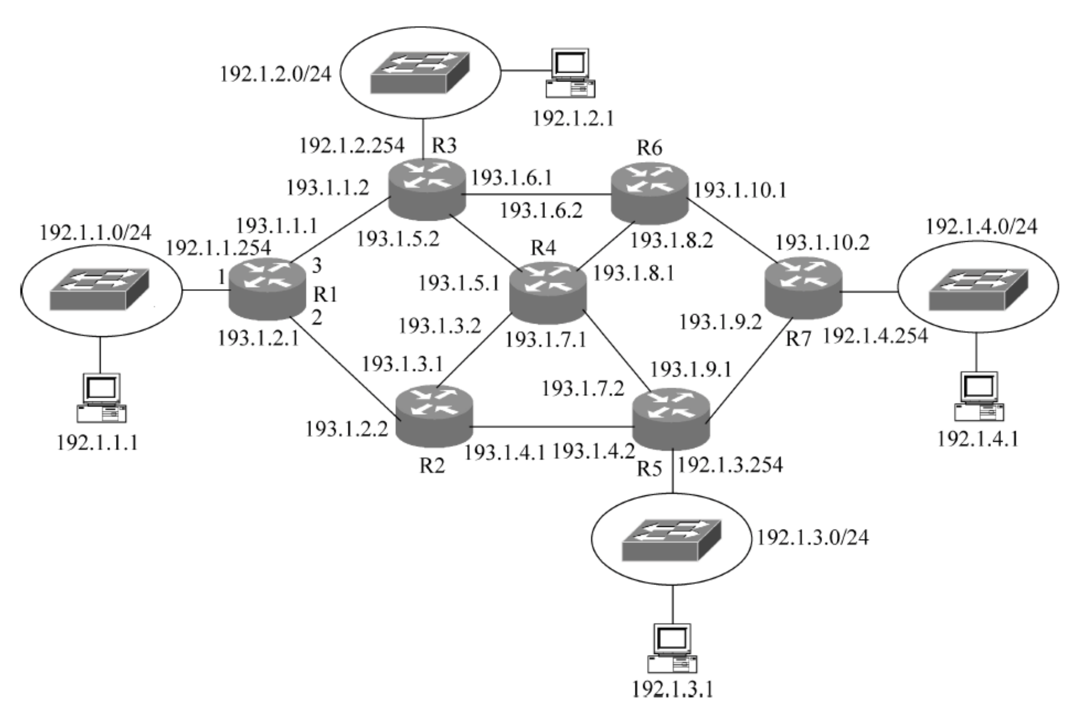


图 6.8　互连网络结构

表 6.17　路由器 R1 直连路由项

目的网络	距离	下一跳路由器
192.1.1.0/24	0	直接

表 6.18　路由器 R3 直连路由项

目的网络	距离	下一跳路由器
192.1.2.0/24	0	直接

表 6.19　路由器 R5 直连路由项

目的网络	距离	下一跳路由器
192.1.3.0/24	0	直接

表 6.20　路由器 R7 直连路由项

目的网络	距离	下一跳路由器
192.1.4.0/24	0	直接

2. 路由器 R1 公告路由消息

路由器 R1 为了让其他路由器获悉通过它可以到达的网络，周期性地公告它所具有的路由项，如＜192.1.1.0/24，0＞，表明经过它可以到达网络 192.1.1.0/24，距离为 0，这些路由项组合成路由消息，路由器 R1 周期性地公告由路由表中全部路由项构成的路由消息。

在本例中,路由器 R1 向它的相邻路由器 R2 和路由器 R3 公告经过它可以到达的网络及距离,如图 6.9 所示。包含这些路由项的路由消息最终封装成 IP 分组,通过路由器 R1 的不同接口发送给相邻路由器,这些 IP 分组的源 IP 地址是路由器 R1 发送该 IP 分组的接口的 IP 地址,由于发送给路由器 R2 和路由器 R3 的 IP 分组从不同的接口发送出去,它们的源 IP 地址并不相同。如果路由器 R1 成了路由器 R2 或路由器 R3 通往某个网络的传输路径上的下一跳路由器,则封装该路由消息的 IP 分组的源 IP 地址就是该路由项的下一跳路由器地址。这些 IP 分组的目的 IP 地址是组播地址 224.0.0.9。路由器 R2 接收到路由器 R1 发送给它的路由消息后,在路由表中添加一项用于指明经路由器 R1 转发后到达网络 192.1.1.0/24 的传输路径的路由项,如表 6.21 所示。同样,路由器 R3 接收到路由器 R1 发送给它的路由消息后,也在路由表中添加用于指明经路由器 R1 转发后到达网络 192.1.1.0/24 的传输路径的路由项,如表 6.22 所示。

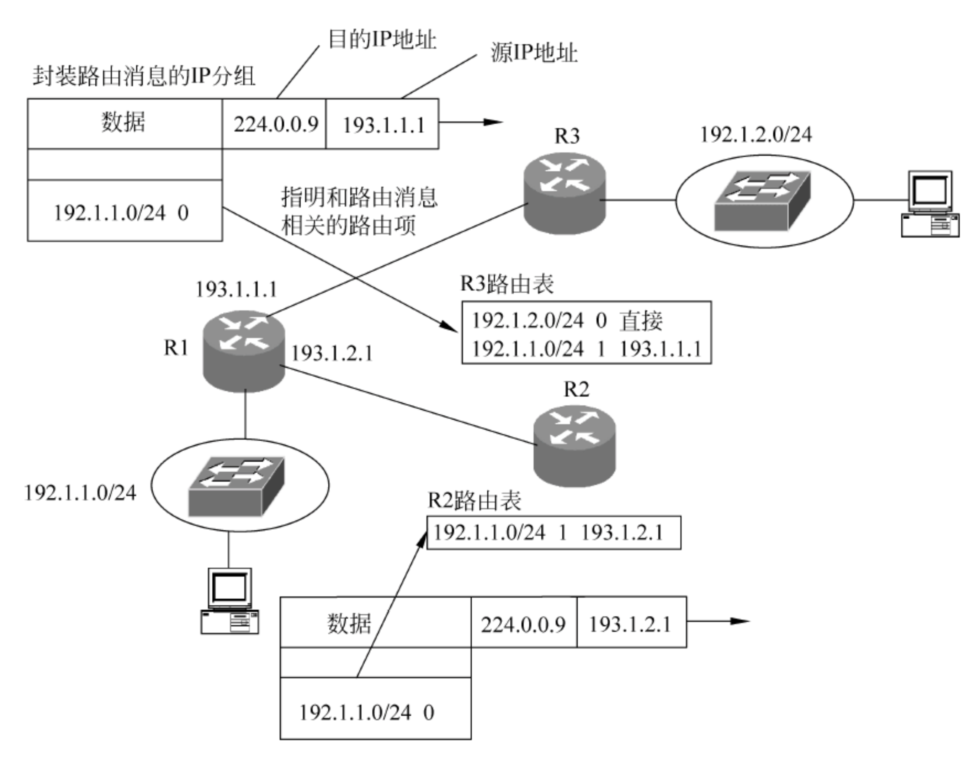


图 6.9 路由器 R1 向路由器 R2 和路由器 R3 公告路由消息的过程

表 6.21 路由器 R2 生成的路由表

目的网络	距离	下一跳路由器
192.1.1.0/24	1	193.1.2.1

表 6.22 路由器 R3 生成的路由表

目的网络	距离	下一跳路由器
192.1.1.0/24	1	193.1.1.1
192.1.2.0/24	0	直接

事实上路由器 R3 也向路由器 R1 公告路由消息，由于本例着重讨论路由器 R5 通过 RIP 建立路由表的过程，因此和路由器 R5 建立路由表无关的操作过程不再赘述。

3. 路由器 R2、路由器 R7 公告路由消息

和路由器 R5 相邻的路由器 R2 和路由器 R7 也周期性地向路由器 R5 公告路由消息，如图 6.10 所示。路由器 R2 公告的路由消息中包含路由项<192.1.1.0/24，1 >，表明经路由器 R2 转发后，能够到达网络 192.1.1.0/24，距离为 1。由于路由器 R5 的路由表中没有用于指明通往网络 192.1.1.0/24 的传输路径的路由项，就在路由表中添加 1 项，如表 6.23 所示。同样，路由器 R7 也向路由器 R5 公告路由消息，路由消息包含路由项<192.1.4.0/24，0>，表明经过路由器 R7 转发后，能够到达网络 192.1.4.0/24，距离为 0，由于路由器 R5 的路由表中没有用于指明通往网络 192.1.4.0/24 的传输路径的路由项，在路由表中添加用于指明通往网络 192.1.4.0/24 的传输路径的路由项，如表 6.23 所示。

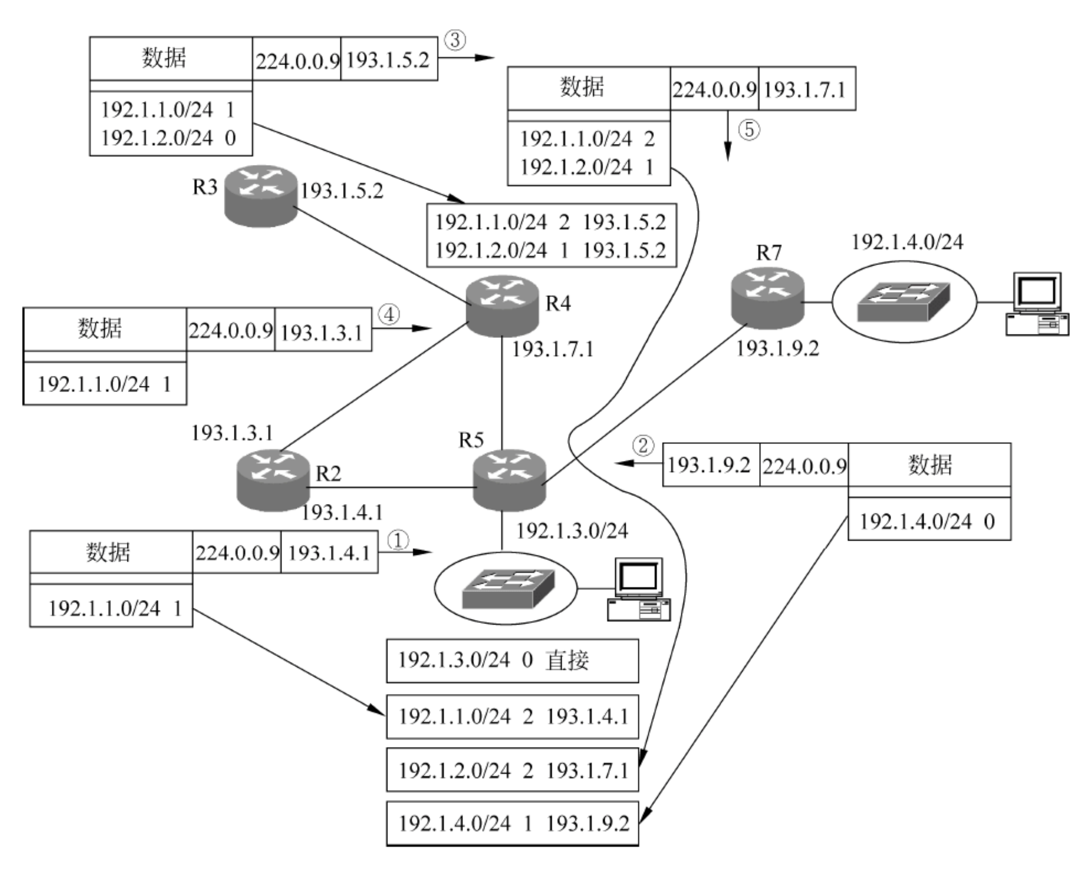


图 6.10 路由器 R5 生成最终路由表的过程

表 6.23 路由器 R5 生成的路由项

目的网络	距离	下一跳路由器
192.1.1.0/24	2	193.1.4.1
192.1.4.0/24	1	193.1.9.2
192.1.3.0/24	0	直接

4. 路由器 R4 接收路由消息

路由器 R4 同样接收到路由器 R2 和路由器 R3 公告给它的路由消息，图 6.10 所示的情况是路由器 R4 先接收到路由器 R3 公告给它的路由消息，在路由表中添加了分别用于指明通往网络 192.1.1.0/24 和 192.1.2.0/24 的传输路径的路由项，如表 6.24 所示。当路由器 R4 接收到路由器 R2 公告给它的路由消息时，发现路由表中已经存在用于指明通往网络 192.1.1.0/24 的传输路径的路由项，根据最短路径原则，路由器 R4 应该选择最短路径作为它的路由项，但在本例中，经路由器 R3 转发后到达网络 192.1.1.0/24 的距离和经路由器 R2 转发后到达网络 192.1.1.0/24 的距离相等。这种情况下，路由器 R4 采用路由表中已有的路由项。反之，如果路由器 R4 先接收到路由器 R2 公告的路由消息，路由器 R4 建立的路由表如表 6.25 所示。

表 6.24 路由器 R4 根据路由器 R3、路由器 R2 的路由消息生成的路由表

目的网络	距离	下一跳路由器
192.1.1.0/24	2	193.1.5.2
192.1.2.0/24	1	193.1.5.2

表 6.25 路由器 R4 根据路由器 R2、路由器 R3 的路由消息生成的路由表

目的网络	距离	下一跳路由器
192.1.1.0/24	2	193.1.3.1
192.1.2.0/24	1	193.1.5.2

5. 路由器 R4 公告路由消息

路由器 R4 也向路由器 R5 公告路由消息，路由消息中包含路由项＜192.1.1.0/24，2＞和＜192.1.2.0/24，1＞，由于路由器 R5 的路由表中没有用于指明通往网络 192.1.2.0/24 的传输路径的路由项，因此，该路由项被添加到路由器 R5 的路由表中，如表 6.26 所示。但路由器 R5 的路由表中已经存在用于指明通往网络 192.1.1.0/24 的传输路径的路由项，而且，该路由项所给出的距离(2)比经过路由器 R4 转发的传输路径所给出的距离(3)小，因此，选择原路由项。路由器 R5 最终生成的路由表如表 6.26 所示，整个过程如图 6.10 所示。

表 6.26 路由器 R5 最终生成的路由项

目的网络	距离	下一跳路由器
192.1.1.0/24	2	193.1.4.1
192.1.4.0/24	1	193.1.9.2
192.1.3.0/24	0	直接
192.1.2.0/24	2	193.1.7.1

分析路由器 R5 通过 RIP 建立路由表的操作过程，可以总结出路由器通过 RIP 生成路由表的步骤：一是经过配置生成到达和其直接相连的网络的路由项；二是通过周期性地和相邻路由器交换各自的路由项，逐渐在所有路由器中建立到达所有网络的路由项。

6.3.4 RIP 动态适应网络变化的过程

RIP 作为路由协议的最大的好处在于能够根据网络拓扑结构的变化，自动调整各个路由器中的路由表。如果图 6.8 所示的网络中路由器 R5 和路由器 R2 之间的通信出现问题，出现通信问题的一种原因可能是连接路由器 R5 和路由器 R2 的物理链路发生故障，这种情况下，路由器 R5 能够立即检测到连接路由器 R2 的物理链路失效，在路由表中删除所有以路由器 R2 为下一跳路由器的路由项(或者将其距离改为无穷大值 16)。另一种原因可能是路由器 R2 发生故障，不再向它的相邻路由器公告路由消息，当然，也不可能正确地转发 IP 分组。这种情况下，路由器 R5 无法立即检测到路由器 R2 的故障，但路由表中的每一项路由项都和定时器相关联，只要从接收到的路由消息中能够重新推导出该路由项，就重新设置一下定时器，因此，只要能够周期性地接收到包含该路由项的路由消息，和该路由项相关联的定时器就不会溢出，该路由项就长期有效。但一旦长时间接收不到包含该路由项的路由消息，就一直无法重新设置和该路由项关联的定时器，最终导致定时器溢出，使该路由项无效。图 6.11 中，路由器 R5 一直接收不到路由器 R2 公告的路由消息，就一直无法重新设置与以路由器 R2 为下一跳路由器的路由项关联的定时器，最终导致定时器溢出，使这些路由项无效。当然无效的结果可以是从路由表中删除该路由项，或将其距离变为无穷大值(16)。一旦以路由器 R2 为下一跳路由器的路由项变为无效，路由器 R5 中就没有用于指明通往网络 192.1.1.0/24 的传输路径的路由项，当接收到路由器 R4 公告的路由消息时，就根据其

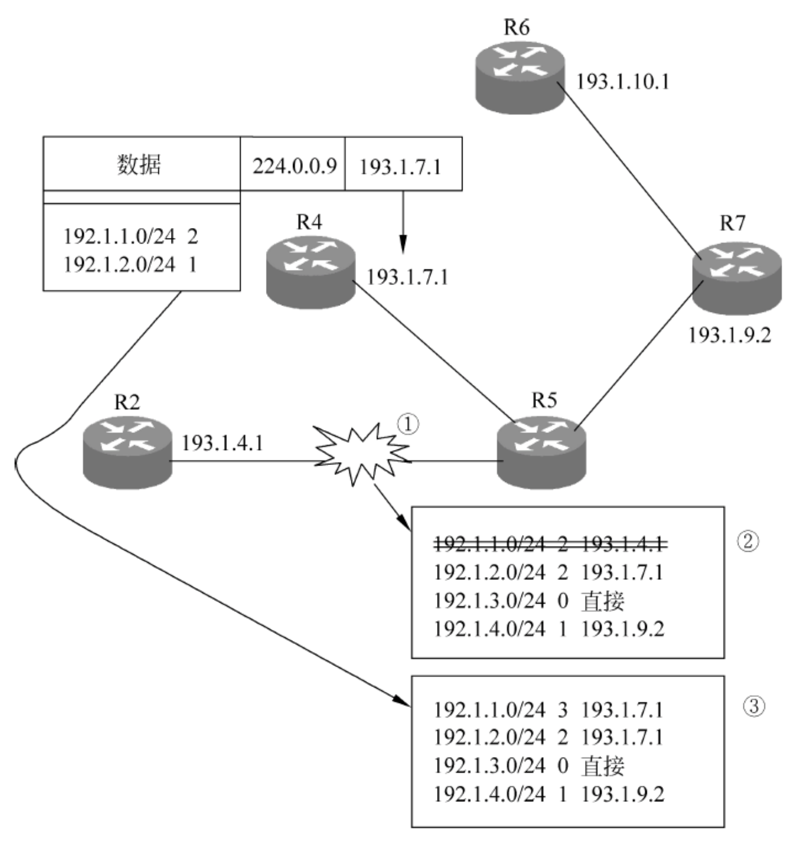


图 6.11 RIP 动态调整路由器 R5 路由表的过程

中包含的和网络 192.1.1.0/24 相关的路由项，推导出以路由器 R4 为下一跳路由器的通往网络 192.1.1.0/24 的传输路径，并将其添加到路由表中，这样，路由器 R5 重新有了用于指明通往网络 192.1.1.0/24 的传输路径的路由项，并以此为根据转发以网络192.1.1.0/24为目的网络的 IP 分组。

6.3.5　计数到无穷大和水平分割

1. 计数到无穷大过程

在图 6.12 (a)所示互连网络正常的情况下，路由器 R1 和路由器 R2 生成如图所示的用于指明通往网络 NET1 的传输路径的路由项。但一旦路由器 R1 连接网络 NET1 的链路发生故障，路由器 R1 中和网络 NET1 关联的路由项的距离将变为 16，表示网络 NET1 不可达。如果

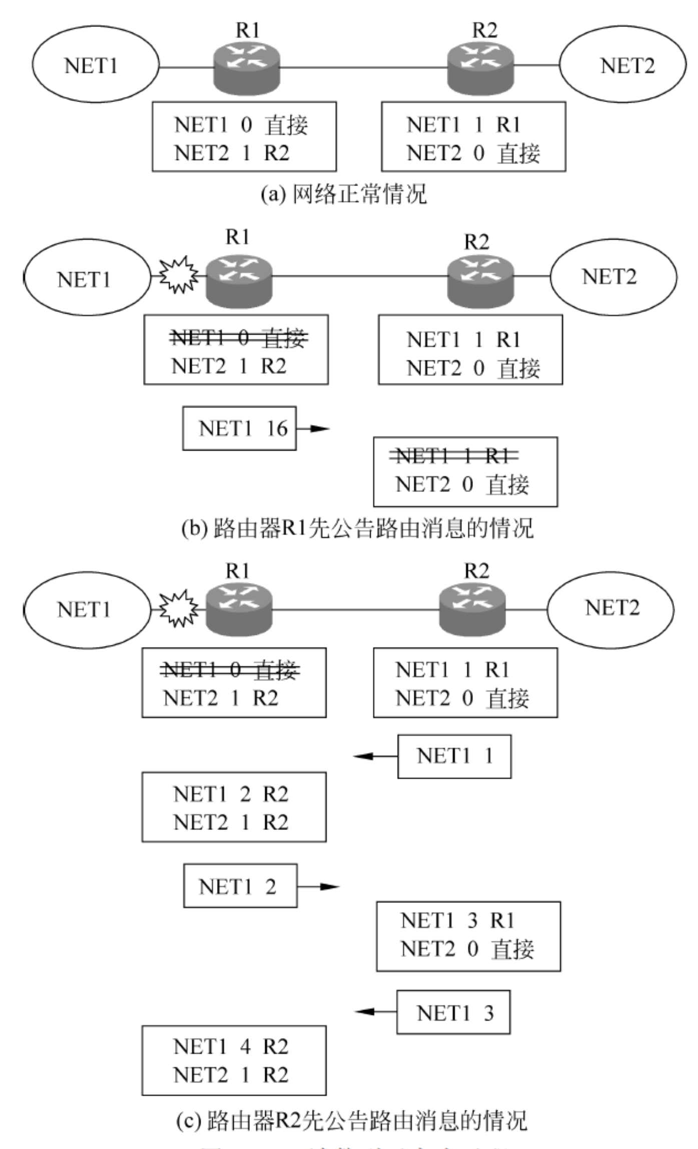


图 6.12　计数到无穷大过程

路由器 R1 先向路由器 R2 发送了和网络 NET1 相关的、距离为 16 的路由项，根据路由器处理路由消息流程中情况⑤的处理方式，路由器 R2 将从路由表中删除目的网络为 NET1、下一跳路由器为 R1 的路由项，路由器 R1、路由器 R2 的路由表趋于稳定，如图 6.12(b)所示。

但如果路由器 R1 在向路由器 R2 发送和网络 NET1 相关的路由项前，先接收了路由器 R2 向它公告的路由消息，通过路由项＜NET1,1＞获悉可以经路由器 R2 转发后，到达网络 NET1，距离为 1。路由器 R1 重新在路由表中生成和网络 NET1 相关的路由项＜NET1,2,R2＞，如图 6.12(c)所示。当然，路由器 R1 也向路由器 R2 公告路由消息，路由消息中包含和网络 NET1 相关的路由项＜NET1,2＞，由于路由器 R2 中和网络 NET1 相关的路由项的下一跳路由器为 R1，因此用新距离 3 代替老距离 1。同样，当路由器 R2 再次向路由器 R1 公告路由消息时，也使路由器 R1 中和网络 NET1 相关的路由项的距离变为 4。经过若干往复，最终使路由器 R1 和路由器 R2 中与网络 NET1 相关的路由项的距离都变成 16，表明网络 NET1 不可达，路由器中路由表趋于稳定，这就是计数到无穷大的问题。

在前面讨论用距离 16 作为网络不可达的标志时已经提出，这样做将极大地限制 RIP 所作用的互连网络的规模，但实际上这是一个无奈的选择，如果上调表示无穷大的值，势必延长图 6.12 (c)所示的计数到无穷大的过程，使互连网络中路由器的路由表一直不能收敛，影响路由器转发 IP 分组的操作。

2. 水平分割

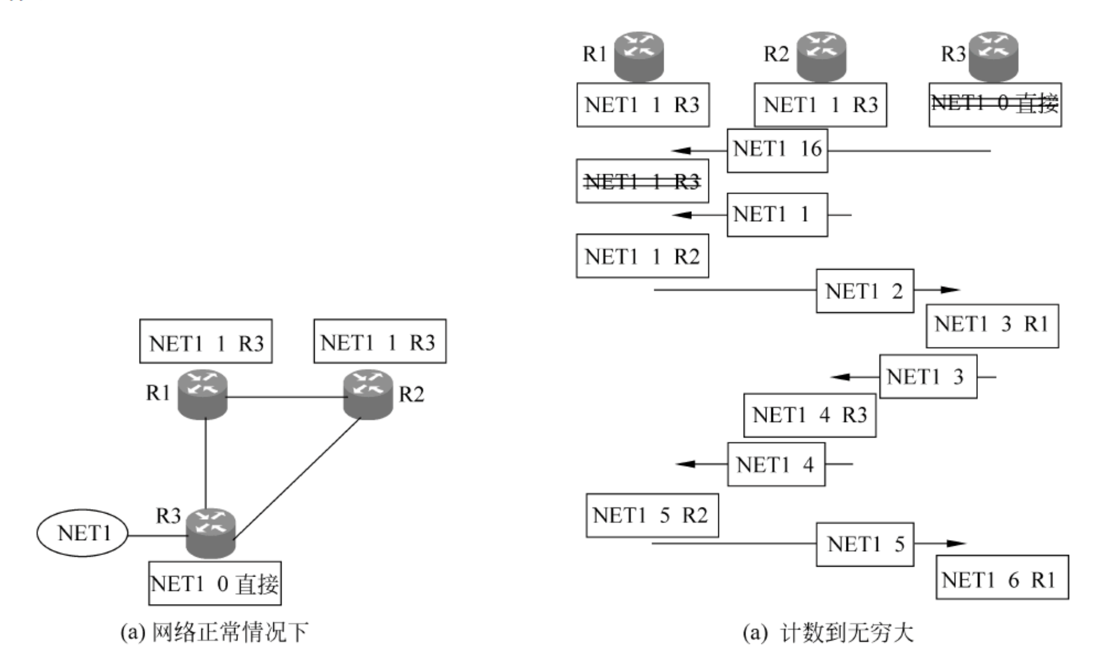
图 6.12(c)所示的计数到无穷大的问题是可以解决的，导致该问题发生的根本原因在于路由器 R2 中和网络 NET1 相关的路由项是通过路由器 R1 公告的路由消息得出的，因此，该路由项的下一跳路由器指明为路由器 R1，而路由器 R2 又向路由器 R1 公告包含该路由项的路由消息，导致该路由项的公告环路(即路由器 R1 对路由器 R2 说经过我转发可以到达网络 NET1，而路由器 R2 又对路由器 R1 说经过我转发可以到达网络 NET1)。消除图 6.12(a)所示互连网络结构下的路由项公告环路问题并不困难，只要规定：如果某个路由项是根据通过某个接口接收到的路由消息得出的，那么，以后从该接口公告的路由消息中不允许包含该路由项，这就是 RIP 的水平分割规则。如果路由器 R2 遵守该规则，那么，通过连接路由器 R1 的接口公告的路由消息中不可能包含以路由器 R1 为下一跳路由器的路由项，图 6.12(c)中的计数到无穷大的问题就不复存在了。

3. 水平分割存在局限

但实际上，即使遵守水平分割规则，计数到无穷大的问题依然可能发生，对于图 6.13(a)所示的互连网络结构，正常情况下，路由器 R1、路由器 R2 和路由器 R3 都能使自己的路由表收敛在一个稳定的状态，如图 6.13(a)所示。一旦路由器 R3 连接网络 NET1 的链路发生故障，路由器 R3 中和网络 NET1 相关的路由项的距离变为 16，表示不可达，如果路由器 R3 能够及时向路由器 R1 和路由器 R2 公告包含路由项＜NET1,16＞的路由消息，路由器 R1 和路由器 R2 将删除和网络 NET1 相关的路由项，所有路由器均认为网络 NET1 不可达。但如果公告路由消息的顺序如下：首先是路由器 R3 向路由器 R1 公告包含路由项＜NET1,16＞的路由消息，导致路由器 R1 中和网络 NET1 相关的路由项被删除。随后，路由器 R2 向路由器 R1 公告路由消息，由于路由器 R2 中和网络 NET1 相关的路由项的下

一跳路由器为路由器 R3，因此，向路由器 R1 公告的路由消息中包含和 NET1 相关的路由项＜NET1，1＞。路由器 R1 根据路由器 R2 向它公告的路由消息推导出和网络 NET1 相关的路由项＜NET1，2，R2＞。这时，路由器 R1 中和网络 NET1 相关的路由项的下一跳路由器为路由器 R2，因此，当路由器 R1 向路由器 R3 公告路由消息时，包含该路由项＜NET1，2＞，使路由器 R3 推导出和网络 NET1 相关的路由项＜NET1，3，R1＞。路由器 R3 同样在向路由器 R2 公告的路由消息中包含该路由项＜NET1，3＞，由于路由器 R2 中和网络 NET1 有关的路由项的下一跳路由器为 R3，用新距离 4 代替老距离 1。如此循环，不断增加和网络 NET1 相关的路由项的距离值，直到无穷大值（16），所有路由器都收敛在网络 NET1 不可达的状态，如图 6.13(b)所示。

RIP 最大的问题就是路由表的收敛过程，在互连网络拓扑结构发生一些变化的情况下，可能需要很长的收敛过程，这一方面由于被迫规定距离值 16 为不可达标志而限制了互连网络规模，另一方面，由于路由表长时间没有收敛在稳定状态而影响了路由器转发 IP 分组的操作。



(a) 网络正常情况下　　(a) 计数到无穷大

图 6.13　计数到无穷大问题

6.4　OSPF

开放最短路径优先（Open Shortest Path First，OSPF）协议是一种链路状态路由协议，OSPF 将路由器每一个接口连接的网络称为链路，路由器通过和相邻路由器交换 Hello 报文确定每一条链路的状态，在确定了所有链路状态后，构建链路状态通告（Link state advertisement，LSA），通过泛洪链路状态通告将自身链路状态通告给互连网络中的所有路由器，每一个路由器在接收到互连网络中所有其他路由器泛洪的链路状态通告后，建立链路状态数据库，链路状态数据库精确描述了互连网络拓扑结构，互连网络中每一个路由器建立

的链路状态数据库是相同的，每一个路由器根据链路状态库构建的以自身为根的最短路径树是一致的。每一个路由器可以根据自身为根的最短路径树构建路由表。

6.4.1 路由器确定自身链路状态

1. Router ID

Router ID 是用于在互连网络中唯一标识某个路由器的路由器标识符，OSPF 可以手工配置 Router ID，也可根据其他配置信息自动生成 Router ID。路由器生成 Router ID 的规则如下。

- 如果为路由器的环路接口（Loopback Interfaces）配置了 IP 地址，用其中值最大的 IP 地址作为该路由器的 Router ID。
- 如果没有为路由器的环路接口配置 IP 地址，在为所有物理接口配置的 IP 地址中选择值最大的 IP 地址作为该路由器的 Router ID。

2. 发现邻居

路由器通过每一个启动 OSPF 的接口周期性地发送 Hello 报文，Hello 报文中包含路由器自身标识符（Router ID），发送该 Hello 报文的接口所属区域的区域标识符，发送该 Hello 报文的接口的子网掩码和优先级，确定的指定路由器标识符和备份指定路由器标识符，邻居列表等。该 Hello 报文被直接封装成 IP 分组，用协议字段值 89 表明净荷是 OSPF 报文，目的 IP 地址是 224.0.0.5，表明该 IP 分组的接收者是网络中所有启动 OSPF 的路由器接口。源 IP 地址是发送该 Hello 报文的接口的 IP 地址。Hello 报文格式和封装过程如图 6.14 所示。Hello 报文的作用是发现邻居，如果两个路由器存在连接在同一个网络上的接口，这两个路由器互为邻居。邻居列表中列出某个路由器已经在该接口所连接的网络上发现的邻居，每一个邻居用其路由器标识符（Router ID）表示。Hello 报文中的发送接口区域标识符、路由器优先级、DR 和 BDR 字段的含义和作用在以后章节讨论。

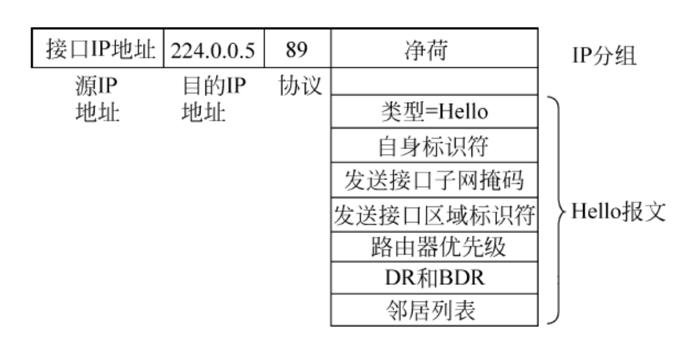


图 6.14 Hello 报文格式和封装过程

图 6.15 是图 6.8 中路由器 R1 和路由器 R2 之间相互发现对方的过程。路由器 R1 将 Hello 报文封装为以接口 IP 地址为源 IP 地址、组播地址 224.0.0.5 为目的 IP 地址的 IP 分组，通过连接路由器 R2 的接口发送出去，组播地址 224.0.0.5 表明接收端是网络内所有启动 OSPF 的其他路由器接口，这里，只有路由器 R2 连接路由器 R1 的接口接收到该 Hello 报文，路由器 R2 根据 IP 分组的源 IP 地址和 Hello 报文中给出的发送接口子网掩码求出发

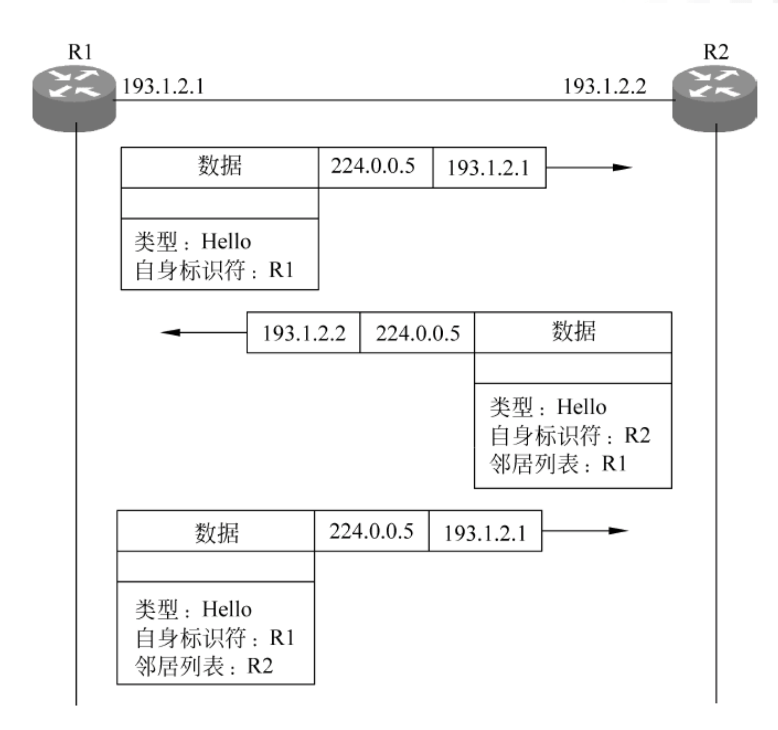


图 6.15 发现邻居过程

送该 Hello 报文的接口的网络地址，同时根据接收该 Hello 报文的接口配置的 IP 地址和子网掩码求出接收该 Hello 报文的接口的网络地址，只有当这两个网络地址相同时，路由器 R2 才继续处理该 Hello 报文，否则，路由器 R2 丢弃该 Hello 报文。路由器 R2 在邻居列表中记录下 Hello 报文中的自身标识符 R1。路由器 R2 发送给路由器 R1 的 Hello 报文中除了自身信息，还需通过邻居列表给出通过该接口接收到的 Hello 报文发现的邻居。当路由器 R1 在路由器 R2 发送给它的邻居列表中发现自身标识符后，确定成功建立和路由器 R2 的邻居关系，同样，当路由器 R2 在随后接收到的路由器 R1 发送的 Hello 报文的邻居列表中发现自身标识符，确定成功建立和路由器 R1 的邻居关系。Hello 报文用于维持和其他路由器的邻居关系，如果某个路由器持续 4×Hello 报文间隔时间没有接收到另一个路由器发送的 Hello 报文，将终止和该路由器之间的邻居关系。

3. 建立邻接关系

某个路由器刚启动时，该路由器中没有其他路由器的链路状态通告，互连网络中的路由器只有在两种情况下泛洪链路状态通告：一是用于指定泛洪链路状态通告周期的定时器溢出；二是某个路由器的链路状态发生改变。某个路由器启动，只会改变和该路由器相邻的路由器的链路状态通告，因此，互连网络中没有和该路由器相邻的其他路由器只有在用于指定泛洪链路状态通告周期的定时器溢出时才会泛洪链路状态通告，为了减少传输开销，路由器在没有链路状态发生改变的情况下，泛洪链路状态通告的周期很长，导致刚启动的路由器需要很长时间才能建立完整的链路状态数据库。为了解决这一问题，要求两个建立邻居关系的路由器必须同步链路状态数据库。两个路由器同步链路状态数据库的过程就是通过发现并下载对方链路状态数据库中存在的、自身链路状态数据库中没有的链路状态通告，使得

两个路由器的链路状态数据库相同的过程。首先介绍建立邻接关系过程中使用的OSPF报文，然后讨论邻接关系建立过程。

1）OSPF报文格式

（1）DD报文格式

数据库描述（Database Description，DD）报文的格式如图6.16所示，用于向对方公告链路状态数据库中存在的LSA，为了减少传输开销，DD中只列出链路状态数据库中存在的LSA的首部。为了保证传输可靠性，采用主从方式，即由主路由器向从路由器发送查询DD报文，从路由器回答应答DD报文，查询和应答DD报文通过序号字段关联在一起，即应答DD报文的序号必须和对应的查询DD报文的序号相同。无论是查询，还是应答DD报文，均可包含用于向对方公告链路状态数据库中存在的LSA的LSA首部列表。

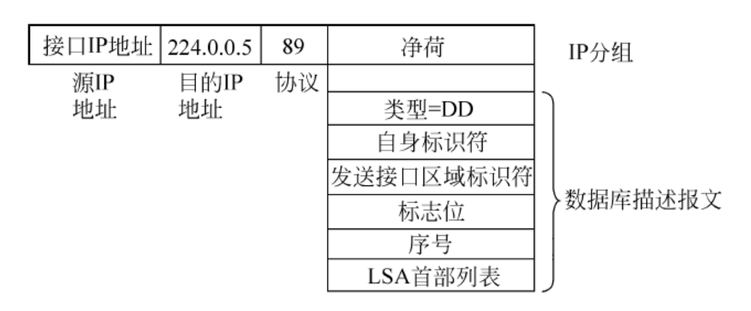


图6.16 数据库描述报文格式

标志位MS用于标识路由器的主从状态，如果MS=1，表示路由器是主路由器，MS=0表示是从路由器，两个路由器中Router ID较大的路由器为主路由器。

标志位I用于标识初始DD报文，对于路由器发送的第一个DD报文，I=1，其他DD报文，I=0。

标志位M是更多DD报文位，如果某个DD报文不是最后一个DD报文，M=1，否则M=0，用M=0的DD报文来表示该次DD报文查询应答过程结束。

（2）LSR报文格式

链路状态请求（Link State Request，LSR）报文格式如图6.17所示，用于请求对方向其传输特定的LSA，用LSA首部列表指定请求传输的LSA。接收到该LSR报文的路由器必须通过链路状态更新（Link State Update，LSU）报文将用LSA首部列表指定的一组完整的LSA传输给LSR发送者。

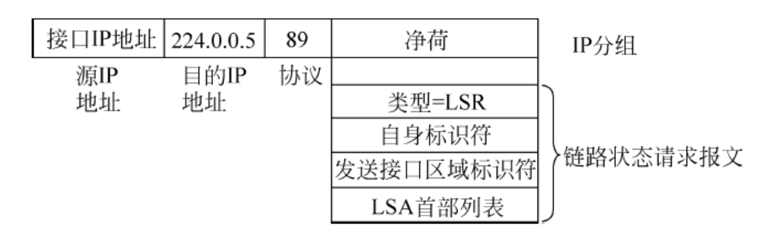


图6.17 链路状态请求报文格式

（3）LSU报文格式

链路状态更新（Link State Update，LSU）报文格式如图6.18所示，它的作用有两个方

面：一是用于向 LSR 发送者传输一组完整的 LSA；二是在路由器自身链路状态发生改变，或是路由器用于指定泛洪链路状态通告周期的定时器溢出时，用于向互连网络中的所有其他路由器泛洪用于表示自身链路状态的 LSA。

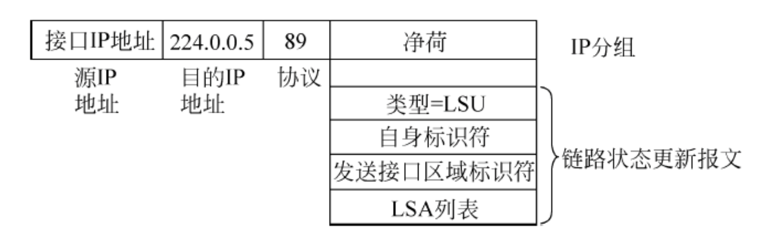


图 6.18　链路状态更新报文格式

2）邻接关系建立过程

邻接关系建立过程如图 6.19 所示。两个路由器建立邻居关系后，才能开始邻接关系建立过程。两个路由器通过交换标志位 I=1 的初始 DD 报文确定主路由器，然后通过反复进行主路由器发送一个查询 DD 报文，从路由器回答一个应答 DD 报文的过程，完成向对方公告链路状态数据库中存在的 LSA 的任务。主路由器每发送一个查询 DD 报文，序号增 1，从路由器回答的应答 DD 报文中的序号必须与对应的查询 DD 报文中的序号相同，最后一个查询 DD 报文和应答 DD 报文的标志位 M=0。如果某个路由器发现对方路由器的链路状态数据库中存在自身链路状态数据库中没有的 LSA，向对方路由器发送 LSR 报文，并在 LSR 报文中用 LSA 首部列表指定需要对方路由器传输的一组 LSA。对方路由器通过 LSU 报文向其传输一组完整的 LSA。一旦两个路由器建立邻接关系，两个路由器的链路状态数据库完成同步过程。

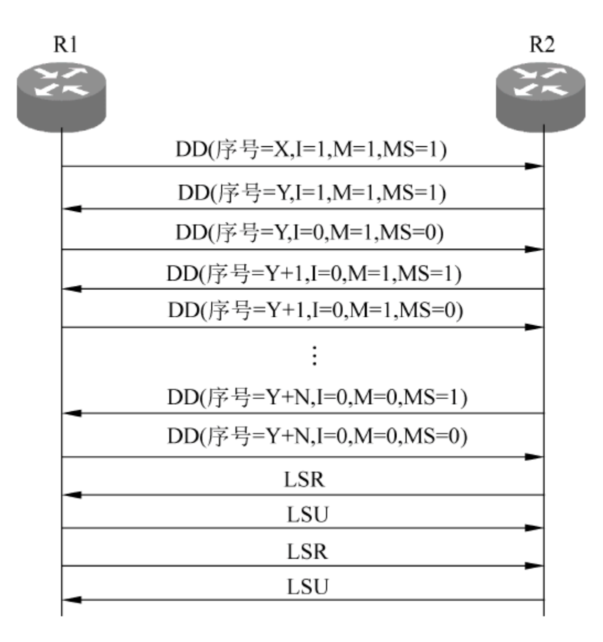


图 6.19　邻接关系建立过程

4. 指定路由器和备份指定路由器

如果某个路由器接口连接的是一个广播型网络(如以太网),该广播型网络上可能同时连接 N 个路由器,如果 N 个路由器两两之间建立邻接关系,需要建立 $N\times(N-1)/2$ 个邻接关系,如图 6.20(b)所示的 4 个路由器之间的 6 个邻接关系。这会大大增加广播型网络的传输开销。为了解决这一问题,在广播型网络中确定一个路由器作为指定路由器(Designated Router,DR),所有其他路由器只和指定路由器建立邻接关系。为了容错,在确定指定路由器的同时,确定一个备份指定路由器(Backup Designated Router,BDR),在指定路由器无法正常工作的情况下,由备份指定路由器取代指定路由器。这样,所有其他路由器只需与指定路由器和备份指定路由器建立邻接关系,当然,指定路由器和备份指定路由器之间也需建立邻接关系,N 个路由器只需建立 $2\times(N-2)+1$ 个邻接关系,如图 6.20(c)所示的 4 个路由器之间的 5 个邻接关系。

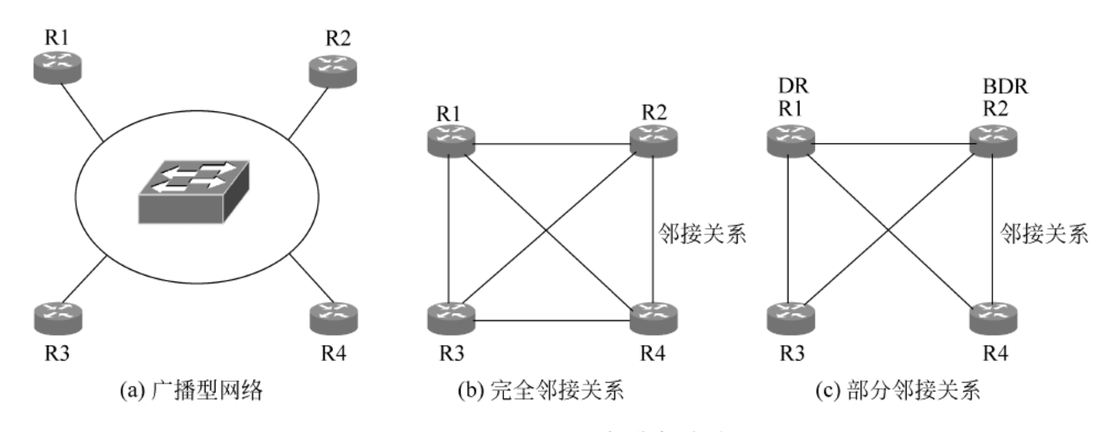


图 6.20 DR、BDR 和部分邻接关系

路由器通过在广播型网络中广播 Hello 报文竞争指定路由器和备份指定路由器,竞争机制可以保证:

- 初始时,在所有建立邻居关系的路由器中选择优先级最高的路由器为指定路由器,如果存在两个以上的具有相同最高优先级的路由器,选择其中 Router ID 最大的路由器为指定路由器,用同样的方式在剩下的路由器中选择备份指定路由器;
- 一旦广播型网络中已经选出指定路由器,即使新接入路由器的优先级大于已经选出的指定路由器的优先级,广播型网络仍然以已经选出的指定路由器为指定路由器;
- 一旦指定路由器和其他路由器终止邻居关系,即使存在优先级大于备份指定路由器的其他路由器,广播型网络仍然选择备份指定路由器为指定路由器,然后,在其他路由器中选出备份指定路由器。

一旦广播型网络选出指定路由器和备份指定路由器,Hello 报文中 DR 和 BDR 字段给出这两个路由器连接广播型网络的接口的 IP 地址。除指定路由器和备份指定路由器以外的其他路由器发送链路状态更新报文时,以 224.0.0.6 为目的 IP 地址,表示接收端是指定路由器和备份指定路由器。

5. 链路状态通告实例

下面以图 6.8 中路由器 R1 为例，讨论路由器建立和其他路由器之间邻接关系后的链路状态。路由器 R1 三个接口分别连接三个网络，其中接口 1 连接末端网络，网络地址和子网掩码分别是 192.1.1.0 和 255.255.255.0，接口 2 和接口 3 连接转接网络，转接网络的主要作用是实现两个路由器互连。假定图 6.8 中所有转接网络都是以太网，序号较大的路由器为该转接网络的指定路由器，接口 2 连接的转接网络中的 DR 为路由器 R2，路由器 R2 连接转接网络的接口的 IP 地址为 193.1.2.2。接口 3 连接的转接网络中的 DR 为路由器 R3，路由器 R3 连接转接网络的接口的 IP 地址为 193.1.1.2。对于末端网络(Link Type=3)，Link ID 和 Link Data 的值分别是网络地址和子网掩码，对于转接网络(Link Type=2)，Link ID 和 Link Data 的值分别是 DR 和始发路由器连接该转接网络的接口的 IP 地址，对于路由器 R1 接口 2 连接的转接网络，Link ID 和 Link Data 的值分别是路由器 R2(DR)和路由器 R1 连接该转接网络的接口的 IP 地址。因此得出表 6.27 所示的路由器 R1 三个接口所连接的链路的链路状态。

表 6.27 图 6.8 中路由器 R1 链路状态

链路类型(Link Type)	链路标识符(Link ID)	链路数据(Link Data)	代价(Cost)
末端网络(3)	192.1.1.0	255.255.255.0	1
转接网络(2)	193.1.2.2	193.1.2.1	1
转接网络(2)	193.1.1.2	193.1.1.1	1

对于图 6.8 所示的互连网络结构，转接网络的作用仅仅是实现两个路由器互连，实际的转接网络可能连接终端，也需生成用于指明通往所有转接网络的传输路径的路由项。因此，路由器 R1 连接的两个转接网络中的 DR 还需生成有关转接网络的链路状态，如表 6.28 所示，链路状态中给出 DR 连接该转接网络的接口的 IP 地址和子网掩码(据此确定转接网络的网络地址)和连接在该转接网络上的所有路由器(Router ID 列表)。

表 6.28 转接网络 DR 链路状态

DR 接口 IP 地址	DR 接口子网掩码	连接路由器
193.1.1.2	255.255.255.252	R1
193.1.2.2	255.255.255.252	R1

针对图 6.8 所示的互连网络结构，本例只讨论用于指明通往末端网络的传输路径的路由项的生成过程，因此，转接网络的链路状态被省略。图 6.21 所示是路由器 R1 根据链路状态生成的 LSA，LSA 首部中给出始发路由器和序号，可以据此唯一确定该 LSA。

6. 例题解析

【例 6.3】 互连网络结构和路由器接口 IP 地址与子网掩码配置如图 6.22 所示，假定路由器 RTA、RTD 和 RTE 分别配置了环路地址 192.168.10.5/32、192.168.10.3/32 和 192.168.10.1/32，填写表 6.29 中的 Router ID 和表 6.30 中的指定路由器。

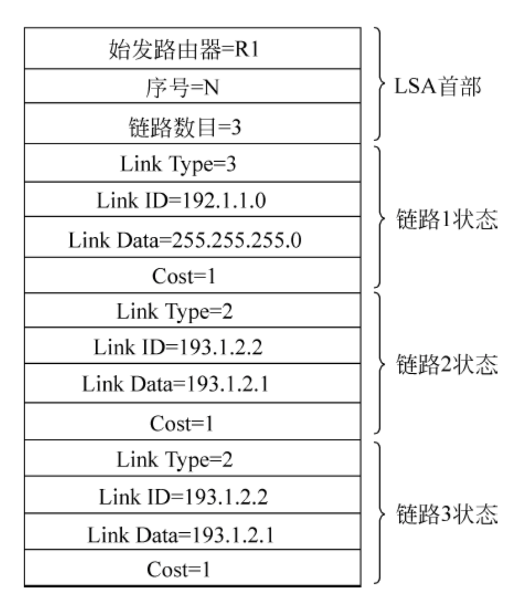


图 6.21 LSA 格式

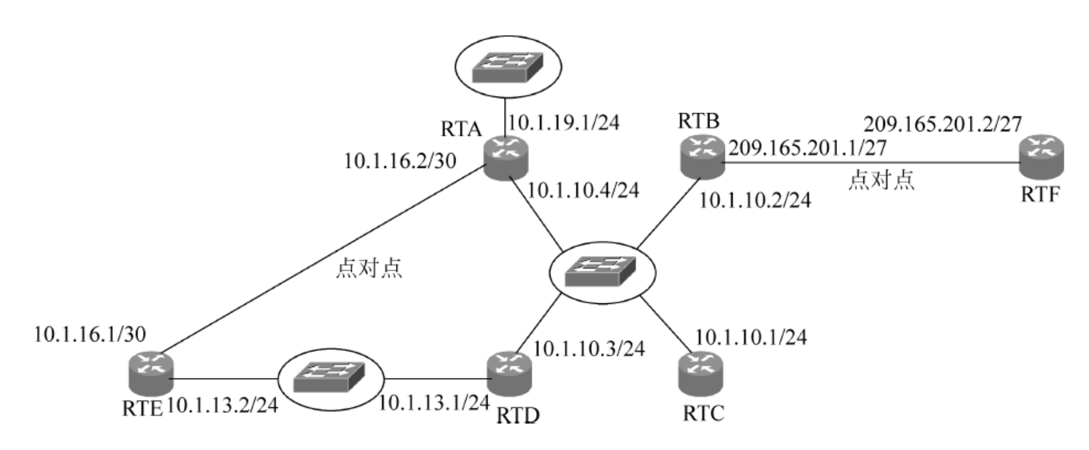


图 6.22 网络结构

表 6.29 各个路由器的 Router ID

路由器	Router ID
RTA	192.168.10.5
RTB	209.165.201.1
RTC	10.1.10.1
RTD	192.168.10.3
RTE	192.168.10.1
RTF	209.165.201.2

表 6.30 各个网络的指定路由器

网络地址	指定路由器
10.1.10.0/24	RTB
10.1.13.0/24	RTD
10.1.16.0/30	不需要
10.1.19.0/24	RTA
209.165.201.0/27	不需要

【解析】 如果为路由器配置了环路接口的 IP 地址，则以环路接口的 IP 地址作为该路由器的 Router ID，否则，以最大的物理接口 IP 地址作为该路由器的 Router ID，因此，路由器 RTA、路由器 RTE 和路由器 RTD 以配置的环路接口 IP 地址作为 Router ID，其他路由器以最大物理接口 IP 地址作为 Router ID。

广播型网络，如以太网，需要产生指定路由器，在连接在广播型网络中的所有路由器中选择 Router ID 最大的路由器作为该广播型网络的指定路由器。网络 10.1.10.0/24 中连接了路由器 RTA、路由器 RTB、路由器 RTC 和路由器 RTD，路由器 RTB 的 Router ID 最大，选择路由器 RTB 为指定路由器。网络 10.1.13.0/24 中连接了路由器 RTD 和路由器 RTE，路由器 RTD 的 Router ID 最大，选择路由器 RTD 为指定路由器。网络 10.1.19.0/24 中只连接了路由器 RTA，以路由器 RTA 为指定路由器。网络 10.1.16.0/32 和网络 209.165.201.0/27 是点对点网络，不需要产生指定路由器。

6.4.2 泛洪链路状态通告

每一个路由器根据自身链路状态构建 LSA 后，用泛洪方式向其他路由器公告 LSA。对于图 6.8 所示的互连网络结构，路由器 R1 用泛洪方式向其他路由器传输 LSA 的过程如图 6.23 所示。路由器 R1 将 LSA 封装成如图 6.24 所示的链路状态更新报文，通过所有启动 OSPF 的接口发送链路状态更新报文，封装链路状态更新报文的 IP 分组的源 IP 地址为发送接口的 IP 地址，目的 IP 地址为组播地址 224.0.0.6，表示接收端是转接网络中的 DR 和 BDR。LSA 中某条链路的状态用＜Link ID，Link Data，Cost＞表示，如路由器 R1 连接路由器 R2 链路的链路状态为＜193.1.2.2，192.1.2.1，1＞，其中 Link ID 是路由器 R2 连接该链路的接口的 IP 地址，Link Data 是路由器 R1 连接该链路的接口的 IP 地址，Cost 是该链路的代价。当某个路由器通过启动 OSPF 的接口接收到链路状态更新报文，用报文中给出的始发路由器标识符和序号比较前面接收到的链路状态更新报文，如果发现前面接收到的链路状态更新报文中存在路由器标识符和当前接收到的链路状态更新报文相同，且序号大于或等于当前接收到的链路状态更新报文的链路状态更新报文，丢弃当前接收到的链路状态更新报文，不再继续转发该链路状态更新报文。否则，存储当前接收到的链路状态更新报文，向发送该链路状态更新报文的路由器发送链路状态确认报文，从除接收该链路状态更新报文的接口以外的所有启动 OSPF 的接口发送该链路状态更新报文。某个启动 OSPF 的接口发送链路状态更新报文前，重新将其封装为 IP 分组，该 IP 分组的源 IP 地址为发送接口的 IP 地址，如果该路由器不是该接口连接的转接网络的 DR，该 IP 分组的目的 IP 地址为组播地址 224.0.0.6，表示接收端是转接网络中的 DR 和 BDR，如果该路由器是该接口连

接的转接网络的 DR,该 IP 分组的目的 IP 地址为组播地址 224.0.0.5,表示接收端是连接在转接网络上的所有启动 OSPF 的接口。经过中间路由器不断转发,路由器 R1 始发的链路状态更新报文遍历互连网络中的所有路由器。链路状态更新报文中的始发路由器和序号用于标识该路由器发送的最新 LSA,因此,路由器发送的不同的 LSA 中的序号是不同的,且随着 LSA 的发送顺序递增,同一路由器发送的 LSA 中,序号最大的 LSA 是最新的。

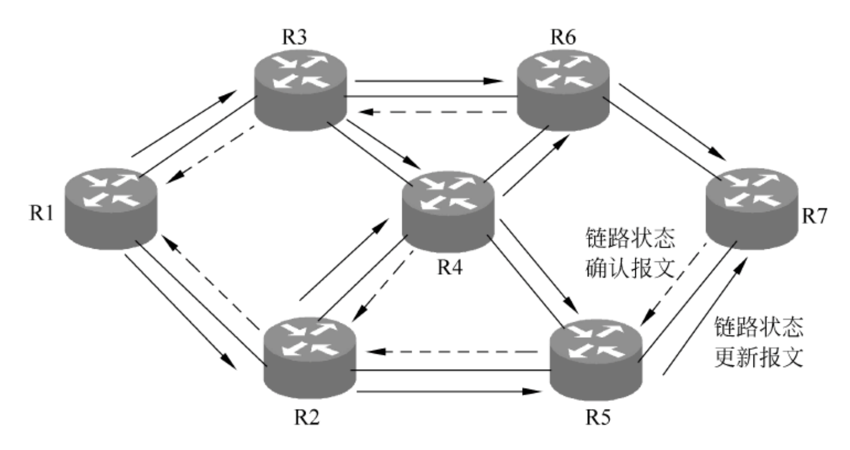


图 6.23　路由器 R1 用泛洪方式传输 LSA 的过程

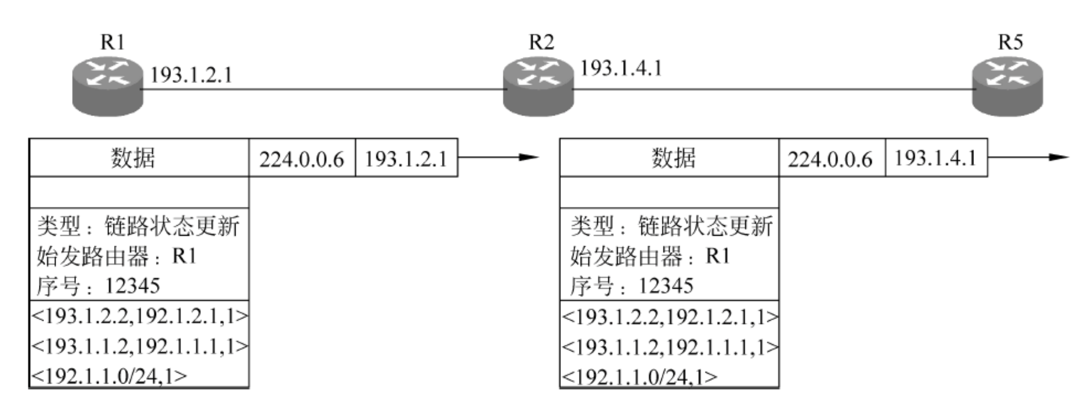


图 6.24　链路状态更新报文内容和封装格式

当所有路由器发送的链路状态更新报文遍历互连网络中所有路由器后,互连网络中每一个路由器都建立了如表 6.31 所示的链路状态数据库。

表 6.31　图 6.8 所示互连网络对应的链路状态数据库

邻居	邻居接口 IP 地址	链路代价
R1 链路状态		
R2	193.1.2.2	1
R3	193.1.1.2	1
192.1.1.0/24		1
R2 链路状态		
R1	193.1.2.1	1
R4	193.1.3.2	1
R5	193.1.4.2	1

续表

邻居	邻居接口 IP 地址	链路代价
	R3 链路状态	
R1	193.1.1.1	1
R4	193.1.5.1	1
R6	193.1.6.2	1
192.1.2.0/24		1
	R4 链路状态	
R2	193.1.3.1	1
R3	193.1.5.2	1
R5	193.1.7.2	1
R6	193.1.8.2	1
	R5 链路状态	
R2	193.1.4.1	1
R4	193.1.7.1	1
R7	193.1.9.2	1
192.1.3.0/24		1
	R6 链路状态	
R3	193.1.6.1	1
R4	193.1.8.1	1
R7	193.1.10.2	1
	R7 链路状态	
R5	193.1.9.1	1
R6	193.1.10.1	1
192.1.4.0/24		1

6.4.3 构建路由表算法

1. 算法描述

当互连网络中所有路由器都构建了表 6.31 所示的链路状态数据库，每个路由器可以计算出到达网络 192.1.1.0/24、192.1.2.0/24、192.1.3.0/24、192.1.4.0/24 的最短路径，并据此构建路由表，但路由表中针对每一个网络的路由项只需给出通往该网络的传输路径上的下一跳路由器，无须给出传输路径经过的所有路由器，对于以特定路由器为根的最短路径树，所有分枝都从该路由器的某个邻居开始，因此，只要求出连接某个网络 N 的分枝的开始路由器 R 和根路由器到达该网络的距离 D，根路由器就可得出该网络对应的路由项<目的网络＝N，距离＝D，下一跳＝R >。根据最短路径算法和每一个网络对应的路由项的特点，得出以下构建路由表中每一个网络对应的路由项的算法。

创建确认列表和临时列表，列表中的每一项是格式为<目的网络，距离，下一跳>的路由项，临时列表中的路由项是中间路由项，确认列表中的路由项是最终路由项。目的网络为根结点的路由项格式为<根结点标识符，0，—>；目的网络为和根结点直接连接

的网络的路由项格式为＜目的网络，链路代价，直接＞；目的网络为和根结点直接相连的路由器的路由项格式为＜路由器标识符，链路代价，路由器标识符＞，对这些和根结点直接相连的路由器，下一跳为自身。如果目的网络为连接在以根结点为树根的最短路径树中某个分枝上的路由器或网络，下一跳为该分枝的开始路由器，即如果某个分枝的开始路由器为根结点相邻路由器 R，则对于连接在该分枝上的所有路由器和网络对应的路由项，下一跳＝R，距离等于从根结点沿着该分枝到达指定路由器或网络所经过的链路的代价之和。

① 初始化确认列表，第 1 项为根结点路由器 S 对应的路由项＜S，0，－＞。

② 假定确认列表中新增的路由项是目的网络为路由器 N 的路由项，初始化时，N＝S，对 N 的所有邻居进行③或④要求的操作。

③ 从链路状态数据库中找出 N 的邻居 R 或直接连接的网络 X，如果 N＝S，则在临时列表中增加路由项＜R，链路代价，R＞或＜X，链路代价，直接＞。

④ 如果 N≠S。距离 D＝目的网络为 N 的路由项中距离＋连接 N 和 R（或 X）的链路代价，下一跳 Y＝目的网络为 N 的路由项中的下一跳，产生路由项＜R，D，Y＞或＜X，D，Y＞。

- 如果确认列表和临时列表中均没有路由项＜R，D，Y＞或＜X，D，Y＞，在临时列表中增加路由项＜R，D，Y＞或＜X，D，Y＞。
- 如果临时列表中存在目的网络为 R 或 X 的路由项，但路由项中的距离大于 D，用路由项＜R，D，Y＞或＜X，D，Y＞取代临时列表中已经存在的目的网络为 R 或 X 的路由项。

⑤ 从临时列表中找出距离最小的路由项，将其移到确认列表，如果临时列表非空，转到②继续处理。

2. 构建路由表举例

表 6.32 给出图 6.8 中路由器 R5 创建路由表的每一个步骤，当路由项＜R2，1，R2＞在步骤 3 被移到确认列表时，需要重新计算和 R2 相邻的路由器或网络相关的路由项，计算结果为路由项＜R5，2，R2＞、＜R4，2，R2＞和＜R1，2，R2＞，由于确认列表中存在目的网络为 R5 的路由项，因此，路由项＜R5，2，R2＞不再增加到临时列表。由于临时列表中存在目的网络为 R4 的路由项＜R4，1，R4＞，且路由项中的距离（1）小于路由项＜R4，2，R2＞中距离（2），因此不能用路由项＜R4，2，R2＞取代临时列表中已经存在的路由项＜R4，1，R4＞。由于确认列表和临时列表中均无路由项＜R1，2，R2＞，将路由项 ＜R1，2，R2＞增加到临时列表。根据最终确认列表中 4 个网络对应的路由项和链路状态数据库中给出的路由器 R2、路由器 R4 和路由器 R7 作为路由器 R5 邻居时的邻居接口 IP 地址，最终生成表 6.33 所示的路由器 R5 路由表。

表 6.32　图 6.8 中路由器 R5 创建路由表过程

步骤	确认列表	临时列表	说　明
1	＜R5，0，—＞		初始化时，确认列表中只有根结点对应的路由项

续表

步骤	确认列表	临时列表	说　明
2	＜R5,0,—＞	＜R2,1,R2＞ ＜R4,1,R4＞ ＜R7,1,R7＞ ＜193.1.3.0/24,1,直接＞	计算和路由器R5直接连接的路由器或网络相关的路由项
3	＜R5,0,—＞ ＜R2,1,R2＞	＜R4,1,R4＞ ＜R7,1,R7＞ ＜193.1.3.0/24,1,直接＞ ＜R1,2,R2＞	将临时列表中距离最小的路由项＜R2,1,R2＞移到确认列表,重新计算和路由器R2相邻的路由器或网络相关的路由项,得到路由项＜R1,2,R2＞
4	＜R5,0,—＞ ＜R2,1,R2＞ ＜R4,1,R4＞	＜R7,1,R7＞ ＜193.1.3.0/24,1,直接＞ ＜R1,2,R2＞ ＜R3,2,R4＞ ＜R6,2,R4＞	将临时列表中距离最小的路由项＜R4,1,R4＞移到确认列表,重新计算和路由器R4相邻的路由器或网络相关的路由项,得到路由项＜R3,2,R4＞和＜R6,2,R4＞
5	＜R5,0,—＞ ＜R2,1,R2＞ ＜R4,1,R4＞ ＜R7,1,R7＞	＜193.1.3.0/24,1,直接＞ ＜R1,2,R2＞ ＜R3,2,R4＞ ＜R6,2,R4＞ ＜193.1.4.0/24,2,R7＞	将临时列表中距离最小的路由项＜R7,1,R7＞移到确认列表,重新计算和路由器R7相邻的路由器或网络相关的路由项,得到路由项＜193.1.4.0/24,2,R7＞
6	＜R5,0,—＞ ＜R2,1,R2＞ ＜R4,1,R4＞ ＜R7,1,R7＞ ＜193.1.3.0/24,1,直接＞	＜R1,2,R2＞ ＜R3,2,R4＞ ＜R6,2,R4＞ ＜193.1.4.0/24,2,R7＞	由于临时列表中距离最小的路由项＜193.1.3.0/24,1,直接＞中的目的网络是末端网络,不会影响其他路由项中的距离值
7	＜R5,0,—＞ ＜R2,1,R2＞ ＜R4,1,R4＞ ＜R7,1,R7＞ ＜193.1.3.0/24,1,直接＞ ＜R1,2,R2＞	＜R3,2,R4＞ ＜R6,2,R4＞ ＜193.1.4.0/24,2,R7＞ ＜193.1.1.0/24,3,R2＞	将临时列表中距离最小的路由项＜R1,2,R2＞移到确认列表,重新计算和路由器R1相邻的路由器或网络相关的路由项,得到路由项＜193.1.1.0/24,3,R2＞
8	＜R5,0,—＞ ＜R2,1,R2＞ ＜R4,1,R4＞ ＜R7,1,R7＞ ＜193.1.3.0/24,1,直接＞ ＜R1,2,R2＞ ＜R3,2,R4＞	＜R6,2,R4＞ ＜193.1.4.0/24,2,R7＞ ＜193.1.1.0/24,3,R2＞ ＜193.1.2.0/24,3,R4＞	将临时列表中距离最小的路由项＜R3,2,R4＞移到确认列表,重新计算和路由器R3相邻的路由器或网络相关的路由项,得到路由项＜193.1.2.0/24,3,R4＞
9	＜R5,0,—＞ ＜R2,1,R2＞ ＜R4,1,R4＞ ＜R7,1,R7＞ ＜193.1.3.0/24,1,直接＞ ＜R1,2,R2＞ ＜R3,2,R4＞ ＜R6,2,R4＞	＜193.1.4.0/24,2,R7＞ ＜193.1.1.0/24,3,R2＞ ＜193.1.2.0/24,3,R4＞	将临时列表中距离最小的路由项＜R6,2,R4＞移到确认列表,重新计算和路由器R6相邻的路由器或网络相关的路由项,没有产生新的或距离更小的路由项

续表

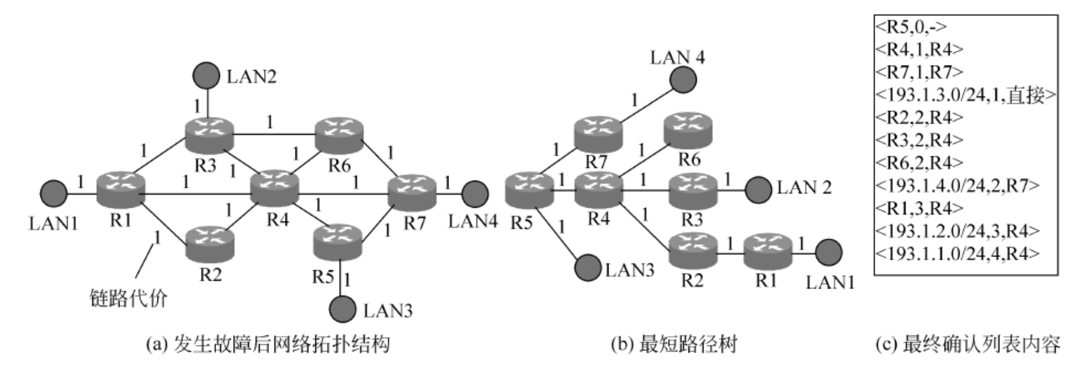
步骤	确认列表	临时列表	说　明
10	＜R5,0,—＞ ＜R2,1,R2＞ ＜R4,1,R4＞ ＜R7,1,R7＞ ＜193.1.3.0/24,1,直接＞ ＜R1,2,R2＞ ＜R3,2,R4＞ ＜R6,2,R4＞ ＜193.1.4.0/24,2,R7＞ ＜193.1.1.0/24,3,R2＞ ＜193.1.2.0/24,3,R4＞		将临时列表中目的网络为末端网络的路由项根据距离大小依次移到确认列表，生成最终确认列表内容

表 6.33　路由器 R5 路由表

目的网络	距离	下一跳路由器
192.1.1.0/24	3	193.1.4.1
192.1.2.0/24	3	193.1.7.1
192.1.3.0/24	1	直接
192.1.4.0/24	2	193.1.9.2

6.4.4　OSPF 动态适应网络变化的过程

如果路由器 R2 和路由器 R5 之间的链路发生故障，或者路由器 R2 或路由器 R5 直接检测到物理连接断开，或者因为长时间无法交换 Hello 报文获知对方不可达。一旦获知对方不可达，路由器 R2 和路由器 R5 将立即通过泛洪标明路由器 R5 和路由器 R2 相互不可达的链路状态更新报文将它们之间相互不可达的信息传播到互连网络中的所有路由器，路由器 R5 最终生成的链路状态数据库将删除路由器 R2 和路由器 R5 互为邻居的链路状态。路由器 R5 根据新的链路状态数据库产生的最短路径树和最终确认列表内容如图 6.25 所示，根据确认列表内容生成的路由表如表 6.34 所示。



(a) 发生故障后网络拓扑结构　(b) 最短路径树　(c) 最终确认列表内容

图 6.25　以路由器 R5 为根的最短路径树

表 6.34　路由器 R5 根据最短路径树生成的路由表

目的网络	距离	下一跳路由器
192.1.1.0/24	4	193.1.7.1
192.1.2.0/24	3	193.1.7.1
192.1.3.0/24	1	直接
192.1.4.0/24	2	193.1.9.2

6.4.5　OSPF 和 RIP 的区别

在 OSPF 中，路由器一旦检测到自身链路状态发生变化，就立即将包含变化后的 LSA 的链路状态更新报文泛洪给互连网络中的所有路由器，而 RIP 是周期性地和相邻路由器交换包含所有路由项的路由消息。因此，OSPF 是将部分信息泛洪给互连网络中所有其他路由器，而 RIP 是将所有信息传输给相邻路由器。在 OSPF 中，每一个路由器可以根据不同的应用要求设定链路代价，也可根据链路状态数据库计算出多条到达指定网络的传输路径，以此实现负载均衡。而 RIP 只能得出最小跳数传输路径。OSPF 由于可以及时更新每一个路由器的链路状态数据库，因此路由表能够及时反映最新的互连网络拓扑结构，而 RIP 存在好消息传得快、坏消息传得慢的问题。

6.4.6　OSPF 分区域建立路由表的过程

RIP 由于存在计数到无穷大的问题，必须用较小的距离值表示无穷大值（RIP 确定为 16），这就使得 RIP 适用的互连网络只能是规模较小的互连网络。OSPF 虽然没有计数到无穷大的问题，但一旦互连网络规模较大，各个路由器泛洪链路状态更新报文造成的传输压力就很大，而且，每一个路由器必须保持和整个互连网络拓扑结构相对应的链路状态数据库，并以此构建路由表。通过表 6.31 已经看到和图 6.8 这样一个小规模互连网络对应的链路状态数据库已经如此复杂，一个大规模互连网络对应的链路状态数据库的复杂程度可想而知，而且根据一个复杂的链路状态数据库来构建路由表的计算过程也十分烦琐、耗时。OSPF 划分区域的功能较好地解决了互连网络规模与链路状态传输开销及构建路由表的计算复杂性之间的矛盾。

1. 划分区域

OSPF 划分区域的方式如图 6.26 所示，整个互连网络被划分成了三个区域：区域 1、区域 2、区域 3，这三个区域都和一个主干区域（区域 0）相连，区域 1 包含路由器 R11、路由器 R12、路由器 R13、路由器 R14 和同时互连区域 1 和主干区域的区域边界路由器 R01、路由器 R02，对于区域边界路由器 R01、路由器 R02，区域 1 称为它们的所在区域。区域 2 包含路由器 R21、路由器 R22、路由器 R23、路由器 R24、路由器 R25 和同时互连区域 2 和主干区域的区域边界路由器 R03、路由器 R04。区域 3 包含路由器 R31、路由器 R32、路由器 R33、路由器 R34 和同时互连区域 3 和主干区域的区域边界路由器 R05、路由器 R06。对于图 6.26 所示的多个区域结构，每一个路由器接口都需要配置该接口所属区域的区域标识符，相邻路由器定义为存在连接在同一个网络且区域标识符相同的接口的路由器，链路状态更新报文中携带始发路由器发送该链路状态更新报文的接口的区域标识符，互连网络中其他路

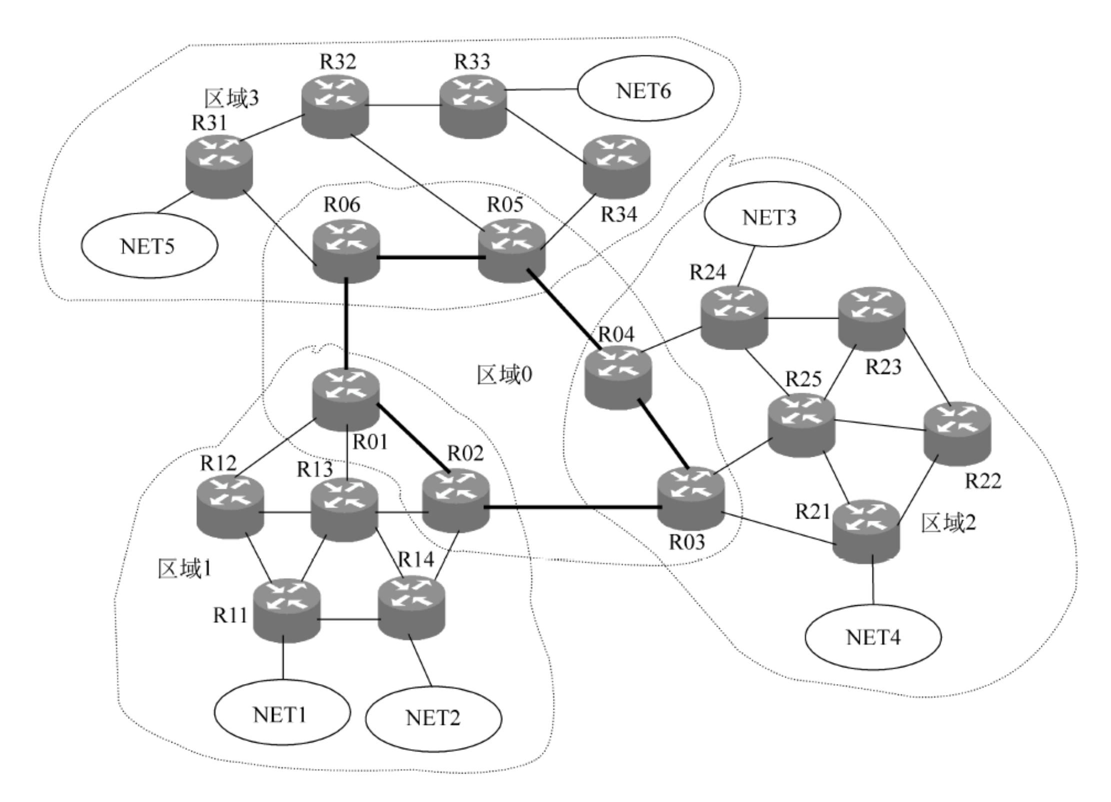


图 6.26 OSPF 划分区域示意图

由器只从配置的区域标识符和该链路状态更新报文携带的区域标识符相同的接口转发该链路状态更新报文。因此,OSPF 只在本区域内作用,即每一个路由器只在本区域内泛洪它的链路状态更新报文,区域内的每一个路由器只记录和本区域的网络拓扑结构相对应的链路状态数据库,并以此为基础构建路由表。那么,某个区域内的路由器如何获知到达另一个区域内网络的传输路径?如区域 1 内的路由器 R11 如何建立用于指明通往网络 NET3、网络 NET4、网络 NET5、网络 NET6 的传输路径的路由项?

2. 建立跨区域传输路径的过程

区域边界路由器同时运行两个分别作用于主干区域和所在区域的 OSPF 进程,如区域边界路由器 R01、路由器 R02,一方面运行作用于区域 1 的 OSPF,最终建立和区域 1 网络拓扑结构相对应的链路状态数据库,并计算出到达区域 1 内网络 NET1、网络 NET2 的传输路径和距离。另一方面又运行作用于主干区域(区域 0)的 OSPF,该 OSPF 在主干区域内泛洪的链路状态更新报文中给出标明主干区域内路由器之间相邻关系的链路状态和到达它所在区域内网络的距离。如路由器 R01 在主干区域内泛洪的链路状态更新报文中不仅给出标明主干区域内路由器之间相邻关系的链路状态,还需给出到达它所在区域内网络 NET1、网络 NET2 的距离。这样一来,主干区域内路由器最终建立的链路状态数据库不仅包含了和主干区域网络拓扑结构相对应的链路状态,还包含了各个区域边界路由器到达其所在区域内网络的距离,根据这样的链路状态数据库所构建的路由表,不仅给出了到达主干区域内其他路由器的最短路径,也给出了到达其他区域内网络的最短路径。同理,区域边界路由器作用于所在区域的 OSPF 在所在区域内泛洪的链路状态更新报文,除了给出标明该区域内路由器之间相邻关系

的链路状态，也需要给出到达其他区域内网络的距离，到达其他区域内网络的距离通过作用于主干区域的OSPF获得。因此，该区域内的路由器最终建立的链路状态数据库，不仅包含了和本区域网络拓扑结构相对应的链路状态，也包含了区域边界路由器到达其他区域内网络的距离，因此，以此为根据构建的路由表能够给出到达其他区域内网络的传输路径和距离。

1）区域边界路由器构建的到达本区域内网络的路由项

每一个区域内的路由器通过和相邻路由器交换Hello报文建立关系，然后向区域内的其他路由器泛洪链路状态更新报文，当区域内所有路由器发送的链路状态更新报文遍历区域内每一个路由器后，区域内每一个路由器建立和本区域网络拓扑结构对应的链路状态数据库，以此为基础构建到达区域内网络的路由项，各个区域边界路由器构建的路由表如表6.35～表6.40所示。

表6.35　路由器R01到达所在区域内网络的路径及代价

目的网络	距离	下一跳路由器
NET1	3	R12
NET2	3	R13

表6.36　路由器R02到达所在区域内网络的路径及代价

目的网络	距离	下一跳路由器
NET1	3	R13
NET2	2	R14

表6.37　路由器R03到达所在区域内网络的路径及代价

目的网络	距离	下一跳路由器
NET3	3	R25
NET4	2	R21

表6.38　路由器R04到达所在区域内网络的路径及代价

目的网络	距离	下一跳路由器
NET3	2	R24
NET4	3	R03

表6.39　路由器R05到达所在区域内网络的路径及代价

目的网络	距离	下一跳路由器
NET5	3	R06
NET6	3	R32

表6.40　路由器R06到达所在区域内网络的路径及代价

目的网络	距离	下一跳路由器
NET5	2	R31
NET6	4	R05

2）主干区域链路状态数据库

区域边界路由器在主干区域内泛洪链路状态更新报文，不仅给出标明主干区域内路由器之间相邻关系的链路状态，还给出到达它所在区域内网络的距离。如路由器 R01 在主干区域内泛洪链路状态更新报文中不仅给出它和路由器 R02、路由器 R06 相邻的链路状态，还给出到达它所在区域内网络 NET1、网络 NET2 的距离。当主干区域内所有路由器发送的链路状态更新报文遍历主干区域内所有路由器后，主干区域内每一个路由器建立表 6.41 所示的链路状态数据库。

表 6.41 主干区域链路状态数据库

邻居或可达网络	链路代价或传输路径距离
R01 链路状态	
R02	1
R06	1
NET1	3
NET2	3
R02 链路状态	
R01	1
R03	1
NET1	3
NET2	2
R03 链路状态	
R02	1
R04	1
NET3	3
NET4	2
R04 链路状态	
R03	1
R05	1
NET3	2
NET4	3
R05 链路状态	
R04	1
R06	1
NET5	3
NET6	3
R06 链路状态	
R05	1
R01	1
NET5	2
NET6	4

3）区域边界路由器构建的到达其他区域内网络的路由项

一旦建立表 6.41 所示的主干区域链路状态数据库，主干区域内的每一个区域边界路由器可以仿照表 6.32 所示的路由器 R5 创建路由表的过程构建路由表，路由表中包含用于指明通往互连网络中所有网络的传输路径的路由项。表 6.42 和表 6.43 分别给出了区域边界路由器 R01、路由器 R02 构建的路由表。当区域边界路由器在所在区域内泛洪链路状态更新报文时，链路状态更新报文中包含其到达其他区域内网络的距离。

表 6.42 区域边界路由器 R01 主干区域路由表

目的网络	距离	下一跳路由器
NET1	3	直接
NET2	3	直接
NET3	5	R02
NET4	4	R02
NET5	3	R06
NET6	5	R06

表 6.43 区域边界路由器 R02 主干区域路由表

目的网络	距离	下一跳路由器
NET1	3	直接
NET2	2	直接
NET3	4	R03
NET4	3	R03
NET5	4	R01
NET6	6	R01

4）区域内路由器构建到达其他区域内网络的路由项

当区域 1 内所有路由器，包括区域边界路由器 R01 和路由器 R02 发送的链路状态更新报文遍历区域 1 内所有路由器后，区域 1 内每一个路由器建立表 6.44 所示的区域 1 内链路状态数据库。当然，区域边界路由器发送的链路状态更新报文，不仅给出标明区域 1 内路由器之间相邻关系的链路状态，还给出其到达其他区域内网络的距离，区域 1 内每一个路由器可以据此计算路由表，例如路由器 R11 计算出的路由表如表 6.45 所示。

表 6.44 区域 1 链路状态数据库

邻居或可达网络	链路代价或传输路径距离
R11 链路状态	
R12	1
R13	1
R14	1
NET1	1

续表

邻居或可达网络	链路代价或传输路径距离
R12 链路状态	
R11	1
R13	1
R01	1
R13 链路状态	
R11	1
R12	1
R14	1
R01	1
R02	1
R14 链路状态	
R11	1
R13	1
R02	1
NET2	1
R01 链路状态	
R12	1
R13	1
R02	1
NET1	3
NET2	3
NET3	5
NET4	4
NET5	3
NET6	5
R02 链路状态	
R01	1
R13	1
R14	1
NET1	3
NET2	2
NET3	4
NET4	3
NET5	4
NET6	6

表 6.45 路由器 R11 路由表

目的网络	距离	下一跳路由器
NET1	1	直接
NET2	2	R14

续表

目的网络	距离	下一跳路由器
NET3	6	R13
NET4	5	R13
NET5	5	R12
NET6	7	R12

5）跨区域传输路径组成

当各个区域内的链路状态数据库稳定后，跨区域传输路径由三段路径组成：一是源区域内路由器至源区域最佳区域边界路由器的传输路径，该传输路径根据源区域的链路状态数据库建立，最佳区域边界路由器是指最短距离的跨区域传输路径经过的区域边界路由器，如路由器 R11 通往 NET6 的传输路径中，源区域最佳区域边界路由器是 R01，路由器 R11 至区域边界路由器 R01 的传输路径 R11→R12→R01 通过区域 1 的链路状态数据库建立；二是源区域最佳区域边界路由器至目的区域最佳区域边界路由器的传输路径，该传输路径根据主干区域的链路状态数据库建立，如路由器 R11 通往 NET6 的传输路径中，源区域最佳区域边界路由器至目的区域最佳区域边界的传输路径是 R01→R06，表 6.42 中路由项＜NET6,5,R06＞反映了这一点；三是目的区域最佳区域边界路由器至目的网络的传输路径，该传输路径根据目的区域的链路状态数据库建立，如路由器 R11 通往 NET6 的传输路径中，目的区域最佳区域边界路由器至目的网络传输路径 R06→R05→R32→R33→NET6 通过区域 3 的链路状态数据库建立。

6.5　BGP

6.5.1　分层路由的原因

图 6.27 是由三个自治系统组成的网络结构，每一个自治系统分配全球唯一的 16 位自治系统号，如 AS1 中的 1，自治系统内部采用内部网关协议，如 RIP 和 OSPF，自治系统之间采用外部网关协议，这里是边界网关协议（Border Gateway Protocol，BGP）。划分自治系统的目的不仅仅是为了解决互连网络规模与路由消息传输开销及计算路由项的计算复杂度之间的矛盾，因为如果将图 6.27 所示的互连网络结构作为单个自治系统，OSPF 可以通过采用划分区域，将链路状态的泛洪范围控制在各个区域内的方法，解决网络规模过大的问题。之所以不能将不同的自治系统作为 OSPF 的不同区域处理，是因为下述原因：一是不同自治系统是由不同管理机构负责管理，因此，很难在代价的取值标准上取得一致，也就很难通过 OSPF 这样的最短路径路由协议求出不同自治系统之间的最佳路由；二是出于安全考虑，自治系统内部结构是不对外公布的，因此，没有人可以在了解各个自治系统的内部结构后，对由多个自治系统组成的互连网络进行区域划分和配置；三是 IP 分组传输过程中选择自治系统时，更多地考虑政策因素和安全因素，这一点和内部网关协议非常不同；四是对于 Internet 这样大规模的网络，用划分区域的方法很难解决互连网络规模与路由消息传输开销及计算路由项的计算复杂度之间的矛盾。因此，自治系统之间需要的是这样一种路由协

议：它可以在不了解各个自治系统内部结构、不需要统一各个自治系统的代价取值标准的情况下，在满足政策和安全的前提下建立自治系统之间的传输路径，而BGP就是这样一种路由协议。

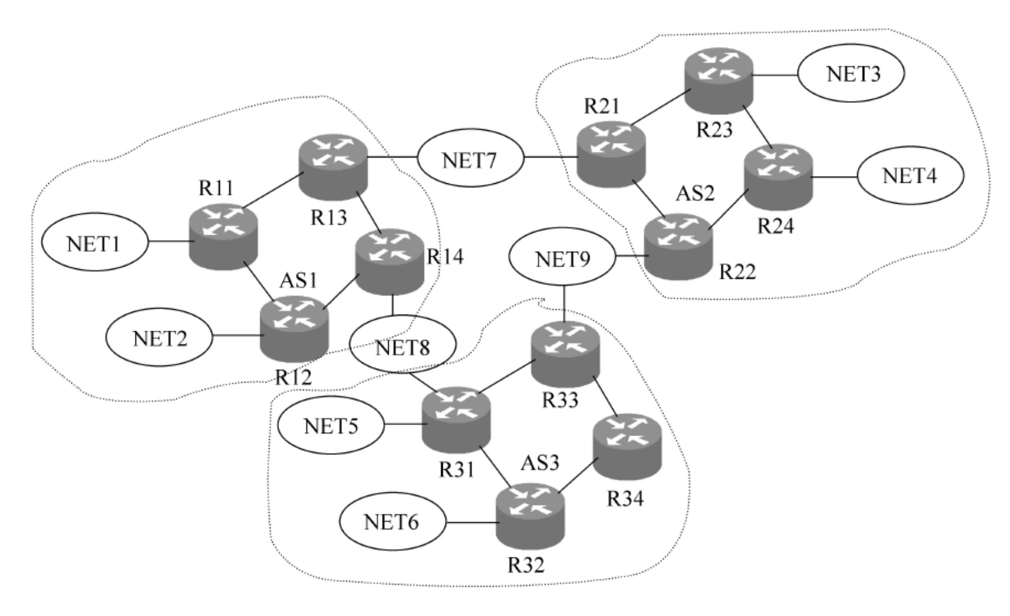


图 6.27 分层路由结构

6.5.2 BGP 报文类型

BGP 定义了 4 种类型的报文，打开(OPEN)报文用于和相邻自治系统中的 BGP 发言人建立邻居关系。保活(KEEPALIVE)报文用于维持和相邻自治系统中的 BGP 发言人之间的邻居关系。更新(UPDATE)报文用于向相邻自治系统中的 BGP 发言人传输路由消息，其中包括新增加的路由和需要撤销的路由。通知(NOTIFICATION)报文用于通知检测到的错误。为了使某个自治系统中的路由器获取到达另一个自治系统中网络的传输路径，自治系统之间需要交换路由消息，为了减少交换路由消息产生的流量，每一个自治系统选择若干路由器作为 BGP 发言人，自治系统之间通过各自的 BGP 发言人交换路由消息。

6.5.3 BGP 工作机制

某个自治系统中，和其他自治系统直接相连的路由器称为自治系统边界路由器，简称为 AS 边界路由器，所谓直接相连是指该路由器和属于另一个自治系统的 AS 边界路由器存在连接在同一个网络上的接口，如图 6.27 中的路由器 R14 和路由器 R31 分别是自治系统 AS1 和自治系统 AS3 的 AS 边界路由器。一般情况下，选择 AS 边界路由器作为 BGP 发言人，两个相邻自治系统的 BGP 发言人往往是两个存在连接在同一个网络上的接口的 AS 边界路由器，如选择路由器 R14 和路由器 R31 分别作为自治系统 AS1 和自治系统 AS3 的 BGP 发言人。每一个 BGP 发言人向其他自治系统中 BGP 发言人发送的路由消息是该自治系统可以到达的网络，以及通往该网络的传输路径经过的自治系统序列，这样的路由消息称为路径向量，如路由器 R31 发送给路由器 R14 的路径向量可以是＜NET5：AS3＞、

＜NET4：AS3，AS2＞，表明经过自治系统 AS3 可以到达网络 NET5，经过自治系统 AS3 和自治系统 AS1 可以到达网络 NET4。对于任何一个特定网络，每一个自治系统选择经过自治系统最少的传输路径作为通往该网络的传输路径。由于 BGP 对任何外部网络，即位于其他自治系统中的网络，选择经过自治系统最少的传输路径作为通往该外部网络的传输路径，因此，称 BGP 为路径向量路由协议。需要注意的是，选择经过自治系统最少的传输路径和选择距离最短的传输路径是不同的，计算距离需要统一度量，而且还需要知道自治系统内部拓扑结构，计算经过的自治系统不需要知道自治系统内部拓扑结构和每一个自治系统对度量的定义。下面通过自治系统 AS1 中路由器 R11 建立通往外部网络的传输路径为例，详细讨论 BGP 工作机制。

1. 建立 BGP 发言人之间的邻居关系

BGP 发言人之间实现单播传输，因此，每一个 BGP 发言人都必须知道和其相邻的 BGP 发言人的 IP 地址。在图 6.27 中，由于需要在自治系统 AS1 中的路由器 R13 和自治系统 AS2 中的路由器 R21、自治系统 AS1 中的路由器 R14 和自治系统 AS3 中的路由器 R31 和自治系统 AS1 中的路由器 R13 和路由器 R14 之间相互交换 BGP 报文，必须在这些 BGP 发言人之间建立邻居关系。为了实现有着邻居关系的两个路由器之间的可靠传输，在通过打开报文建立这两个路由器之间的邻居关系前，须先建立这两个路由器之间的 TCP 连接，以此保证 BGP 报文的可靠传输。

2. 自治系统各自建立内部路由

每一个自治系统通过各自的内部网关协议建立到达自治系统内各个网络的传输路径，表 6.46、表 6.47 和表 6.48 给出了自治系统 AS1 中路由器 R11，自治系统 AS2 和自治系统 AS3 中 BGP 发言人（AS 边界路由器 R21、路由器 R31）通过内部网关协议建立的用于指明到达自治系统内各个网络的传输路径的路由项。

表 6.46 路由器 R11 路由表

目的网络	距离	下一跳路由器
NET1	1	直接
NET2	2	R12
NET7	2	R13
NET8	3	R12

表 6.47 路由器 R21 路由表

目的网络	距离	下一跳路由器
NET3	2	R23
NET4	3	R23
NET7	1	直接
NET9	2	R22

表 6.48 路由器 R31 路由表

目的网络	距离	下一跳路由器
NET5	1	直接
NET6	2	R32
NET8	1	直接
NET9	2	R33

3. BGP 发言人之间交换路由信息

如图 6.28 所示，建立邻居关系的 BGP 发言人之间相互交换更新报文，更新报文中给出通过它所在的自治系统能够到达的网络，通往这些网络的传输路径经过的自治系统序列及下一跳路由器地址，如果交换更新报文的两个 BGP 发言人属于不同的自治系统，如路由器 R13 和路由器 R21，下一跳路由器地址给出的是 BGP 发言人发送更新报文的接口的 IP 地址，而这一接口通常和相邻自治系统的 BGP 发言人的其中一个接口连接在同一个网络上。如果交换更新报文的两个 BGP 发言人属于同一个自治系统，如路由器 R13 和路由器 R14，下一跳路由器地址是原始更新报文中给出的地址，本例中，路由器 R13 转发的来自路由器 R21 的更新报文中的下一跳路由器地址仍然是路由器 R21 连接网络 NET7 的接口的 IP 地址，图 6.28(c)中用路由器 R21 表示。当自治系统 AS1 中 BGP 发言人接收过相邻自治系统中 BGP 发言人发送的更新报文，同时又在自治系统 AS1 中 BGP 发言人之间交换过各自接收到的更新报文后，自治系统 AS1 中 BGP 发言人建立如表 6.49 所示的用于指明通往外部网络的传输路径的路由项，路由类型 E 表明目的网络位于其他自治系统。

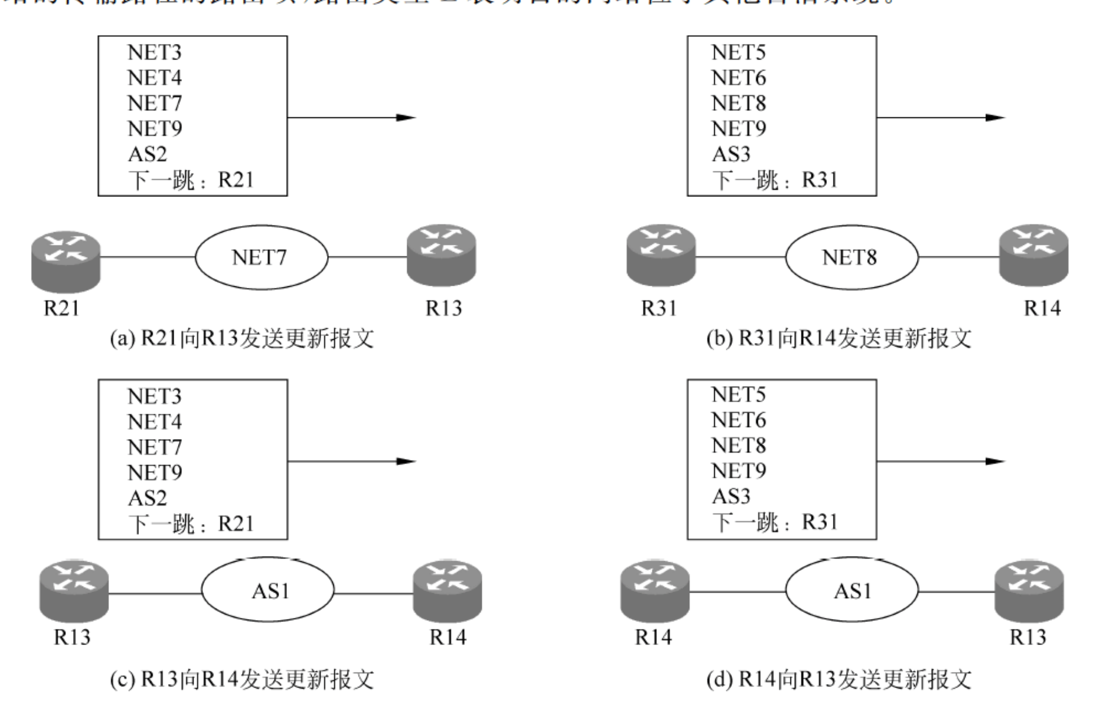


图 6.28 相邻 BGP 发言人相互交换更新报文的过程

表 6.49 AS1 中 BGP 发言人建立的对应外部网络的路由项

目的网络	距离	下一跳路由器	路由类型	经历的自治系统
NET3		R21	E	AS2
NET4		R21	E	AS2
NET5		R31	E	AS3
NET6		R31	E	AS3
NET9		R21	E	AS2

表 6.49 中路由项＜NET3，R21，AS2＞中下一跳路由器 R21 的作用是用于给出通往自治系统 AS2 的传输路径，为了建立自治系统 AS1 通往自治系统 AS2 的传输路径，当自治系统 AS2 中路由器 R21 向自治系统 AS1 中的 BGP 发言人路由器 R13 发送路径向量时，还需给出自己连接网络 NET7 的接口的 IP 地址。注意：NET7 是互连路由器 R13 和路由器 R21 的网络，它既和自治系统 AS1 相连，又和自治系统 AS2 相连，由于自治系统 AS1 内部网关协议建立的路由表包含了用于指明通往属于自治系统 AS1 的所有网络的传输路径的路由项，自然包含目的网络为 NET7 的路由项，因此，在确定路由器 R21 连接网络 NET7 的接口的 IP 地址为自治系统 AS1 通往自治系统 AS2 传输路径上的下一跳 IP 地址后，能够结合自治系统 AS1 内部网关协议建立的路由表创建用于指明通往网络 NET3 的传输路径的路由项。

实际 BGP 操作过程中，所有建立相邻关系的 BGP 发言人之间不断交换更新报文，然后由 BGP 发言人选择经过的自治系统最少的传输路径作为通往某个外部网络的传输路径，并记录在路由表中。由于本例只讨论路由器 R11 建立完整路由表过程，故和该过程无关的更新报文交换过程不再赘述。

4. 路由器 R11 建立完整路由表过程

路由器 R11 通过内部网关协议建立表 6.46 所示的用于指明通往属于本自治系统的所有网络的传输路径的路由项，在本自治系统中的 BGP 发言人建立表 6.49 所示的目的网络为外部网络的路由项后，通过内部网关协议向本自治系统中的其他路由器公告表 6.49 所示的路由项，当路由器 R11 接收到本自治系统中的 BGP 发言人路由器 R13 或路由器 R14 公告的表 6.49 所示的目的网络为外部网络的路由项后，结合表 6.46 所示的目的网络为内部网络（属于本自治系统的网络）的路由项，得出表 6.50 所示的完整的路由表，其中目的网络为外部网络的路由项中给出的下一跳是路由器 R11 通往表 6.49 中给出的下一跳路由器的自治系统内传输路径上的下一跳路由器，如表 6.49 中目的网络为 NET3 的路由项中的下一跳是路由器 R21，实际表示的是路由器 R21 连接 NET7 的接口的 IP 地址，路由器 R11 通往 NET7 的传输路径上的下一跳是路由器 R13，距离是 2，因此，通往外部网络 NET3 的本自治系统内传输路径上的下一跳是路由器 R13，距离是 2。需要指出的是，自治系统中的 BGP 发言人选择通往外部网络的传输路径时，选择的依据是经过的自治系统最少的传输路径。自治系统内的其他路由器只是被动接受本自治系统中的 BGP 发言人选择的通往外部网络的传输路径，然后根据内部网关协议生成的路由项确定自治系统内通往外部网络的这一段传输路径，无论是路由项中的距离，还是下一跳路由器都是对应这一段传输路径的，这一段传输路径实际上是路由器通往本自治系统连接相邻自治系统的网络的传输路径，而该

相邻自治系统是通往该外部网络的传输路径经过的第一个自治系统。

表 6.50 路由器 R11 完整路由表

目的网络	距离	下一跳路由器	路由类型	经历的自治系统
NET1	1	直接	I	
NET2	2	R12	I	
NET3	2	R13	E	AS2
NET4	2	R13	E	AS2
NET5	3	R12	E	AS3
NET6	3	R12	E	AS3
NET7	2	R13	I	
NET8	3	R12	I	
NET9	2	R13	E	AS2

习题

6.1 为什么路由协议得出的端到端传输路径是由一系列路由器组成的？路由表中的下一跳路由器和当前路由器之间有什么限制？

6.2 为什么说 RIP 是好消息传得快，坏消息传得慢？根据图 6.8 所示互连网络举例说明。

6.3 什么是 RIP 的计数到无穷大的问题？能否彻底解决？

6.4 根据 RIP 操作过程，求出图 6.8 中路由器 R3 路由表的收敛过程。

6.5 RIP 的水平分割有什么作用？

6.6 RIP 为距离设置无穷大值的原因是什么？对 RIP 造成什么限制？

6.7 假定路由器 B 的路由表如表 6.51 所示，现路由器 B 接收到路由器 C 发来的如表 6.52 所示的路由消息，试求出路由器 B 更新后的路由表(详细说明每一个步骤)。

表 6.51 路由器 B 路由表

目的网络	距离	下一跳
N1	7	A
N2	2	C
N6	8	F
N8	4	E
N9	4	F

表 6.52 路由器 C 发送的路由消息

目的网络	距离
N2	4
N3	8
N6	4
N8	3
N9	5

6.8 假定互连网络中结点 A 和结点 F 的路由表如表 6.53 和表 6.54 所示,距离为跳数,画出和这两个结点路由表一致的互连网络拓扑结构图。

表 6.53 结点 A 路由表

结点	距离	下一跳结点
B	1	B
C	1	C
D	2	B
E	3	C
F	2	C

表 6.54 结点 F 路由表

结点	距离	下一跳结点
A	2	C
B	3	C
C	1	C
D	2	C
E	1	E

6.9 OSPF 优于 RIP 的地方是什么?

6.10 根据 OSPF 工作原理,给出求出图 6.29 中结点 B 至其他各个结点最短路径的步骤。

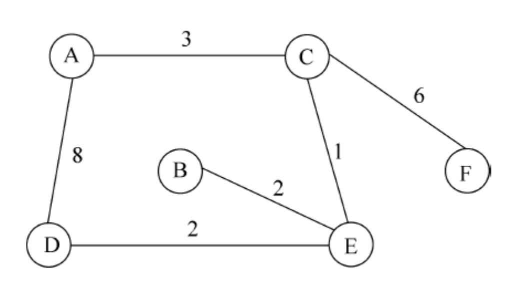


图 6.29 题 6.10 图

6.11 OSPF 如何保证只在本区域内泛洪链路状态?

6.12 为什么 OSPF 需要划分区域?

6.13 根据 OSPF 操作过程,求出图 6.26 中路由器 R31 路由表的收敛过程。

6.14 为什么需要 BGP? 不能用 OSPF 取代 BGP 的原因是什么?

6.15 OSPF 得出的到达其他区域中网络的传输路径是最短路径吗? 解释原因。

6.16 为什么说 BGP 是路径向量协议,它和 RIP 的最大不同是什么?

6.17 BGP 得出的到达其他自治系统中网络的传输路径是最短路径吗? 解释原因。

6.18 根据 BGP 操作过程,求出图 6.27 中路由器 R12 的路由表收敛过程。

第7章 组播

视频点播、远程教学和网络电视等新型业务要求实现点对多点数据通信,实现点对多点数据通信需要一种只需发送单个分组就能实现源终端至分布在多个不同网络、属于同一组播组的一组终端的分组传输的技术,这种分组传输技术就是组播技术。

7.1 组播基本概念

7.1.1 组播与单播和广播的区别

1. 组播定义

组播是一种实现点对多点通信的分组传输技术,通过组播,源终端只需发送单个分组就能完成向分布在多个不同网络、属于同一组播组的一组终端传输分组的过程。

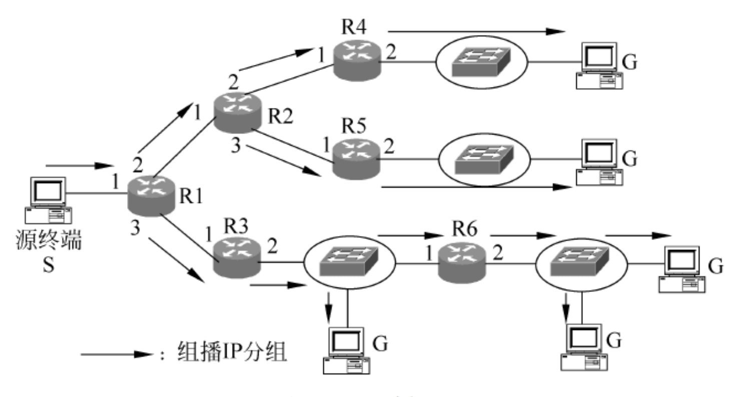


图 7.1 组播过程

组播过程如图 7.1 所示,一组分布在多个不同的网络,且属于同一组播组的终端用唯一的组播地址标识,源终端传输给属于该组播组的一组终端的 IP 分组用唯一标识该组播组的组播地址作为该 IP 分组的目的 IP 地址,这样的 IP 分组称为组播 IP 分组。源终端只发送单个组播 IP 分组,路由器在必要的分枝复制该组播 IP 分组,组播 IP 分组沿着源终端至属于该组播组的一组终端的最短路径到达属于该组播组的所有终端。

2. 单播、广播和组播

以单播地址作为目的 IP 地址的 IP 分组称为单播 IP 分组,单播实现点对点通信,源终

端发送的 IP 分组沿着源终端至目的终端的最短路径到达目的终端。

以广播地址作为目的 IP 地址的 IP 分组称为广播 IP 分组，存在两种类型的广播地址：一是 32 位全 1 的受限广播地址，以这种广播地址为目的 IP 地址的 IP 分组只能在本地网络中广播，即路由器不转发此类 IP 分组；二是网络号字段为特定网络号，主机号字段全 1 的直接广播地址，以这种广播地址为目的 IP 地址的 IP 分组在由网络号指定的特定网络内广播，除了直接连接该特定网络的路由器，其他路由器像转发单播 IP 分组一样转发该广播 IP 分组。

单播和组播的最大不同在于 IP 分组的目的终端，单播 IP 分组的目的终端是由 IP 单播地址指定的唯一终端。组播 IP 分组的目的终端是由组播地址指定的一组终端。

广播和组播的最大不同在于目的终端的分布范围，广播 IP 分组的目的终端是连接在某个特定网络上的所有终端，或本地网络上除源终端以外的所有其他终端。目的终端必须连接在同一个网络上，即广播只在单个网络内进行，而且是连接在特定网络上的所有终端。组播 IP 分组的目的终端可以分布在多个不同的网络上，而且对于连接在任何特定网络上的终端，其中的任何一部分终端均可成为目的终端。

对于路由器，单播 IP 分组和以直接广播地址为目的 IP 地址的广播 IP 分组的间接交付过程是相同的，不同的是直接交付过程。对于单播 IP 分组，路由器通过地址解析过程获取目的终端唯一的链路层地址（如以太网 MAC 地址），将 IP 分组封装成以该链路层地址为目的地址的链路层帧，通过互连路由器和目的终端的网络实现该链路层帧路由器至目的终端的传输过程。对于广播 IP 分组，该 IP 分组被封装成以链路层广播地址（如以太网全 1 MAC 地址）为目的地址的链路层帧，将该链路层帧广播给连接在特定网络中的所有终端。

对于组播，一是路由器必须了解属于某个特定组播组的一组终端的分布范围，即那些网络连接了属于该特定组播组的终端；二是通过组播路由表建立源终端至所有连接了属于某个特定组播组的终端的网络的传输路径。由于可能有多个网络连接了属于该特定组播组的终端，某个路由器可能有多个接口连接多条通往这些网络的传输路径，该路由器必须通过连接这些传输路径的多个接口输出以唯一标识该特定组播组的组播地址为目的地址的组播 IP 分组。如图 7.1 所示，由于多个网络连接了属于组播组 G 的终端，路由器 R1 存在多条分别通往这些网络的传输路径，并由接口 2 和接口 3 分别连接这些传输路径，当路由器 R1 接收到以唯一标识组播组 G 的组播地址为目的 IP 地址的组播 IP 分组时，路由器 R1 必须同时通过接口 2 和接口 3 输出该组播 IP 分组。

7.1.2 组播地址

1. 组播地址格式

组播地址格式如图 7.2 所示，如果 IPv4 32 位 IP 地址的最高 4 位为 1110，表示是组播地址，余下 28 位用于标识组播组。组播地址范围为 224.0.0.0～239.255.255.255，其中 224.0.0.0～224.0.0.255 为预留组播地址，用于特定用途，这些组播地址也称为著名组播地址。224.0.1.0～238.255.255.255 可用于标识用户组播组。239.0.0.0～239.255.255.255 作为本地管理组播地址。下面就是一些常用的著名组播地址，这些组播地址表明接收端是同一网络内的特定结点。

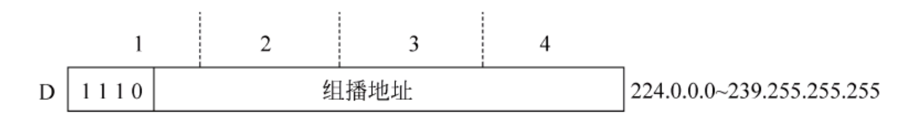


图 7.2 组播地址格式

224.0.0.1 表示网络中所有支持组播的终端和路由器。
224.0.0.2 表示网络中所有支持组播的路由器。
224.0.0.4 DVMRP 路由器。
224.0.0.5 表示网络中所有运行 OSPF 进程的路由器。
224.0.0.6 表示 OSPF 中的指定路由器(DR)。
224.0.0.9 表示网络中所有运行 RIP 进程的路由器。
224.0.0.13 表示网络中所有运行 PIM 的路由器。
224.0.0.18 表示网络中运行 VRRP 进程的路由器。

2. 组播组与组播地址之间关联

每一个组播组由组播地址唯一标识,但组播组与组播地址之间如何建立关联?目前存在两种建立组播组与组播地址之间关联的机制,一是手工配置,通过为特定应用(如某个课程的远程教学)相关的组播组手工配置组播地址建立该组播组与组播地址之间的关联。二是通过动态分配协议动态建立组播组与组播地址之间的关联,动态分配协议有会话目录工具(Session Directory Tool,SDT)、组播地址动态客户端分配协议(Multicast Address Dynamic Client Allocation Protocol,MADCAP)等。由于目前组播技术还没有在互联网范围得到广泛应用,大部分应用都由内部网络实现,因此,目前主要通过手工配置建立组播组与组播地址之间的关联。

每一个终端接收 IP 分组前,需建立接收列表,终端只接收目的 IP 地址在接收列表中的 IP 分组。在为终端配置 IP 地址后,接收列表中列出为终端配置的 IP 地址、32 位全 1 的受限广播地址和根据为终端配置的 IP 地址与子网掩码求出的直接广播地址,如果终端希望加入某个组播组,必须为该组播组分配组播地址,并将该组播地址加入终端的接收列表。

需要指出的是,组播组 G 与分配给组播组 G 的 IP 组播地址是不同的,发送给属于组播组 G 的终端的 IP 分组应该以分配给组播组 G 的 IP 组播地址为目的 IP 地址,为了表述简单,不再区分组播组 G 和分配给组播组 G 的 IP 组播地址,可以直接用组播地址 G 表示分配给组播组 G 的 IP 组播地址。

7.1.3 组播实现技术

1. 构建组播树

互连网络结构如图 7.3 所示,假定互连网络中除网络 192.1.1.0/24 外,所有末端网络都连接属于组播组 G 的终端,源终端 S 至所有连接属于组播组 G 的终端的末端网络的最短路径如图 7.4 所示。图 7.4 所示的以源终端为根的最短路径树称为(S,G)组播树,(S,G)组播树的作用就是将源终端 S 发送的单个目的 IP 地址为组播地址 G 的组播 IP 分组传输到

所有属于组播组 G 的终端,为了构建图 7.4 所属的(S,G)组播树,需要完成以下任务。

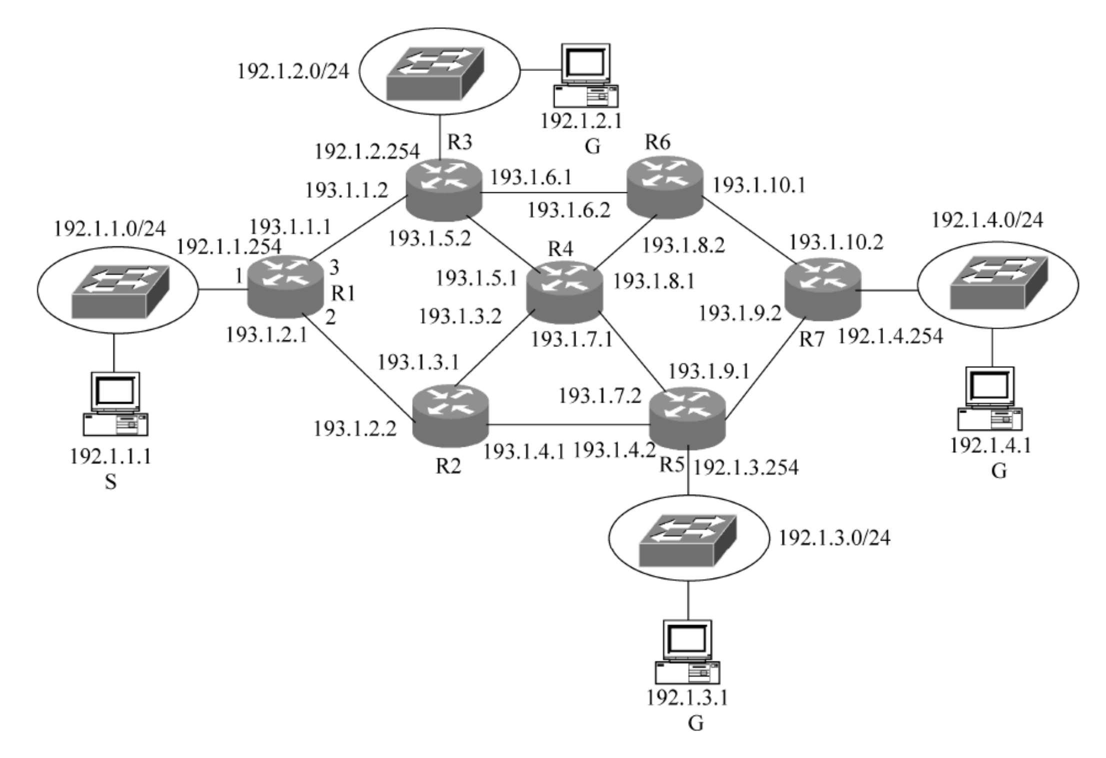


图 7.3　互连网络结构

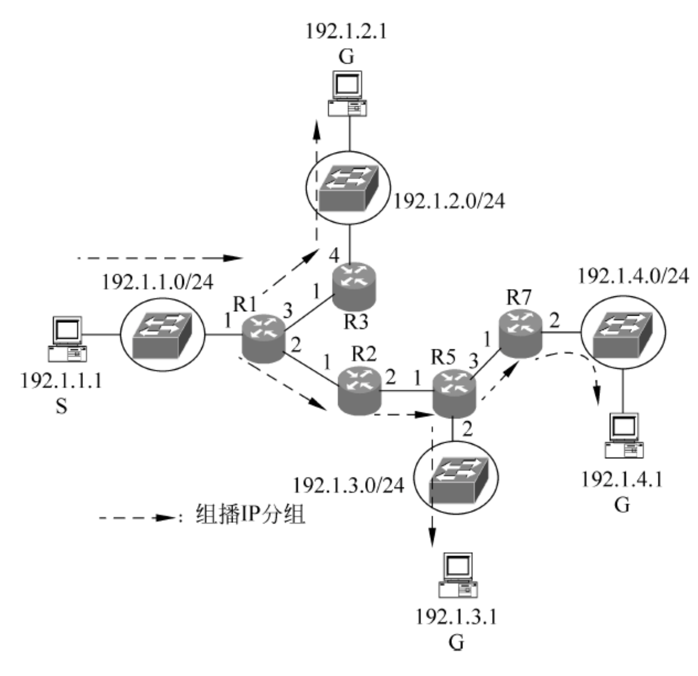


图 7.4　(S,G)组播树

1）构建组播路由表

反映图 7.4 所示的(S,G)组播树的组播路由表需要针对每一对源终端 S 和组播组 G 的组合(S,G)给出上游接口和下游接口列表,上游接口连接通往源终端 S 的最短路径,下游接

口列表中的每一个接口连接一条通往某个连接属于组播组 G 的终端的末端网络的最短路径,对于路由器 R1 和路由器 R2,其组播路由表如表 7.1 和表 7.2 所示。

表 7.1 路由器 R1 组播路由表

源终端	组播组	上游接口	下游接口列表
192.1.1.1/32	G	1	2,3

表 7.2 路由器 R2 组播路由表

源终端	组播组	上游接口	下游接口列表
192.1.1.1/32	G	1	2

2）确定属于组播组 G 的终端的分布范围

属于组播组 G 的终端可以分布在多个不同的网络中,图 7.4 所示的组播树只包含源终端 S 通往连接属于组播组 G 的终端的网络的最短路径,因此,每一个路由器首先需要确定在直接连接的网络中哪些网络连接属于组播组 G 的终端,然后将这些信息扩散到互连网络中的其他路由器,使得每一个路由器的组播路由表中对应(S,G)对的下游接口列表中列出的每一个接口都连接一条通往某个连接属于组播组 G 的终端的末端网络的最短路径。

3）构建交换机组播表

以太网 MAC 地址分为单播地址、广播地址和组地址三种,单播地址用于唯一标识某个以太网终端,交换机通过地址学习过程在转发表中建立单播地址和输出端口之间的关联,当交换机通过某个端口接收到以单播地址为目的 MAC 地址的 MAC 帧时,通过转发表中与该目的 MAC 地址关联的输出端口输出该 MAC 帧。64 位全 1 的 MAC 地址作为广播地址,当交换机通过某个端口接收到以广播地址为目的 MAC 地址的 MAC 帧时,通过除接收端口以外的所有其他端口输出该 MAC 帧。组播过程如图 7.5 所示,交换式以太网实现组播需要解决两个问题:一是需要建立 IP 组播地址至 MAC 组地址的映射规则,二是由于交换机需要通过一组端口输出以 MAC 组地址为目的地址的 MAC 帧,交换机必须建立如图 7.5 所示的组播表,组播表中的每一项是 MAC 组地址与一组输出端口之间建立的关联,当交换机接收到以某个 MAC 组地址为目的地址的 MAC 帧时,交换机通过组播表中与该 MAC 组地址关联的一组输出端口输出该 MAC 帧。

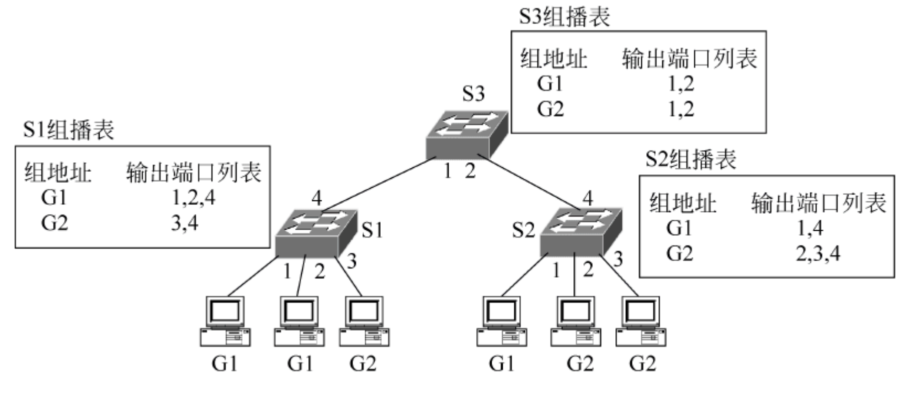


图 7.5 交换机组播表

对于图 7.5 所示的属于组播组 G1 和组播组 G2 的终端的分布情况，首先将 IP 组播地址 G1 和 IP 组播地址 G2 映射到 MAC 组地址，然后在交换机中建立组播表，对于每一个 MAC 组地址，确定以该 MAC 组地址为目的地址的 MAC 帧的一组输出端口。如图 7.5 中交换机 S1 组播表所示，IP 组播地址 G1 对应的 MAC 组地址所关联的一组输出端口是端口 1、端口 2 和端口 4，当交换机接收到以该 MAC 组地址为目的地址的 MAC 帧时，将该 MAC 帧通过端口 1、端口 2 和端口 4 输出。当然，如果该 MAC 帧通过输出端口列表列出的某个端口输入，该 MAC 帧将通过输出端口列表中除接收该 MAC 帧以外的其他所有端口输出该 MAC 帧。

2. 组播相关构件和协议

实现组播需要终端、交换机和路由器协调合作，对于图 7.3 中的末端网络，如网络 192.1.1.0/24，终端、交换机及直接连接末端网络的路由器之间通过互联网组管理协议(Internet Group Management Protocol，IGMP)完成如下功能：

- 让路由器了解属于特定组播组的终端的分布情况；
- 交换机建立组播表；
- 允许终端动态加入或离开某个组播组。

路由器之间通过组播路由协议建立组播路由表，组播路由协议通常分为两类。一类组播路由协议独立于单播路由协议，但建立组播路由表时需要通过单播路由协议建立的单播路由表确定上游接口和下游接口列表，链路代价、最短路径的定义与建立单播路由表的单播路由协议一致。另一类组播路由协议自己确定源终端至连接属于特定组播组的终端的多个网络的最短路径，以此为基础构建组播路由表，链路代价和最短路径含义由组播路由协议自定义。前一类组播路由协议有协议无关组播——稀疏方式(Protocol Independent Multicast-Sparse Mode，PIM-SM)和协议无关组播——密集方式(Protocol Independent Multicast-Dense Mode，PIM-DM)，后一类组播路由协议有距离向量组播路由协议(Distance Vector Multicast Routing Protocol，DVMRP)。

7.2 IGMP

IGMP 有三个版本，分别是 IGMPv1、IGMPv2 和 IGMPv3，目前常用的是 IGMPv2，IGMPv3 在 IGMPv2 的基础上增加了目的终端选择源终端的能力，对于特定组播组，目的终端可以只接收特定源终端发送的组播 IP 分组，或是拒绝接收特定源终端发送的组播 IP 分组，这里主要讨论 IGMPv2。

7.2.1 IGMP 消息类型和格式

如图 7.6 所示，IGMP 消息被直接封装为 IP 分组，用协议字段值 2 表示 IP 分组净荷是 IGMP 消息。

IGMP 消息分为路由器发送的查询消息和终端发送的报告消息与离开消息。查询消息又分为普遍查询消息和指定组播组查询消息。普遍查询消息用于路由器确定直接连接的网

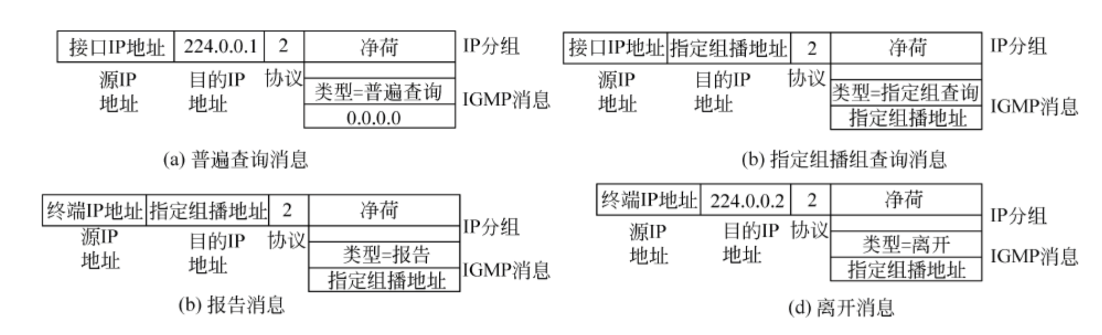


图 7.6 IGMP 消息格式

络中是否有终端加入了组播组，普遍查询消息中的组播地址字段值为 0，封装普遍查询消息的 IP 分组的源 IP 地址是发送该消息的路由器接口的 IP 地址，目的 IP 地址是组播地址 224.0.0.1，表明接收端是所有运行 IGMP 的终端和路由器。指定组播组查询消息用于路由器确定直接连接的网络中是否有终端加入了指定组播组，指定组播组查询消息中的组播地址字段值为标识指定组播组的组播地址，封装普遍查询消息的 IP 分组的源 IP 地址是发送该消息的路由器接口的 IP 地址，目的 IP 地址是标识指定组播组的组播地址，因此，只有加入了指定组播组的终端才会接收和处理该 IGMP 消息。

报告消息用于终端向路由器报告终端加入组播组的情况，终端在两种情况下发送报告消息。一是接收到路由器发送的查询消息(包括普遍查询消息和指定组播组查询消息)后，作为响应消息，向路由器发送报告消息。二是终端新加入某个组播组时，通过向路由器发送报告消息告知路由器直接连接的网络中有终端加入某个组播组。封装报告消息的 IP 分组的源 IP 地址是为终端配置的 IP 地址，目的 IP 地址是标识终端加入的组播组的组播地址，报告消息中的组播地址字段给出标识终端加入的组播组的组播地址，如果终端同时加入多个组播组，接收到普遍查询消息后，需要发送多个报告消息，每一个报告消息中指出终端加入的某个组播组。

7.2.2 IGMP 操作过程

1. 竞争查询者

如果某个网络连接了多个路由器，一是组播路由协议必须保证只把需要在该网络中组播的组播 IP 分组传输给其中一个路由器，否则有可能发生重复传输组播 IP 分组的情况。二是多个路由器中，只需要一个路由器作为查询者，发送查询消息。如图 7.7 所示网络结构中，路由器 R1 和路由器 R2 都和交换式以太网互连，在没有启动 IGMP 侦听功能前，交换式以太网以广播方式传输以 MAC 组地址为目的地址的 MAC 帧，因此任何终端和路由器发送的以 MAC 组地址为目的地址的 MAC 帧被连接在交换式以太网上的所有其他终端和路由器接收。路由器 R1 和路由器 R2 初始时，将自己作为查询者，启动定时器，周期性地发送普遍查询消息。当某个路由器接收到普遍查询消息后，将封装该普遍查询消息的 IP 分组的源 IP 地址与接收该普遍查询消息的接口的 IP 地址比较，如果接口的 IP 地址小于封装该普遍查询消息的 IP 分组的源 IP 地址，维持自己查询者身份不变，周期性发送普遍查询消息。如果接口的 IP 地址大于封装该普遍查询消息的 IP 分组的源 IP 地址，禁止发送查询消息，

启动禁止定时器，每接收到其他路由器发送的查询消息，刷新禁止定时器，只要禁止定时器不溢出，该路由器一直禁止发送查询消息。某个网络的查询者也称为该网络的指定路由器(Designated router，DR)。

2. 查询和报告过程

查询者周期性发送普遍查询消息，每一个终端接收到普遍查询消息后，分别为自己加入的每一个组播组设置报告定时器，报告定时器的初值是某个范围内的随机值，一旦某个组播组关联的报告定时器溢出，该终端发送一个报告消息，报告消息中给出标识该组播组的组播地址，并以该组播地址作为封装该报告消息的 IP 分组的目的 IP 地址。如果在某个组播组关联的报告定时器溢出前，接收到其他终端发送的用于向路由器报告加入该组播组情况的报告消息，该终端将关闭报告定时器，不再发送用于向路由器报告加入该组播组情况的报告消息。路由器接收到报告消息后，在连接该网络的接口中记录网络中终端加入组播组的情况，对于图 7.7 所示组播组分布情况，路由器 R1 和路由器 R2 接口 1 记录的终端加入组播组的情况如表 7.3 所示。

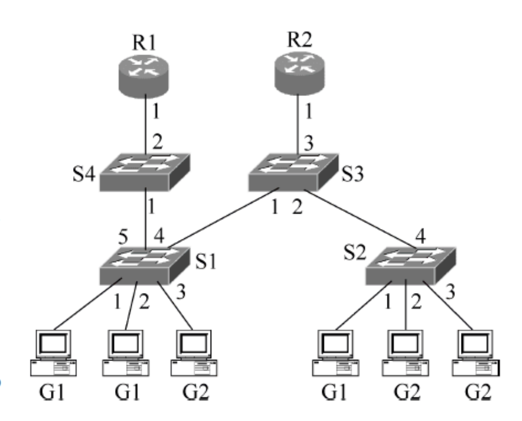


图 7.7 终端加入组播组情况

表 7.3 路由器 R1 和路由器 R2 记录的组播组情况

路由器接口	接口连接的网络中终端加入组播组的情况
路由器 R1 接口 1	G1、G2
路由器 R2 接口 2	G1、G2

当某个终端新加入某个组播组，该终端将立即发送用于向路由器报告加入该组播组情况的报告消息。

路由器接口记录的每一个组播组都关联一个定时器，每当通过该接口接收到终端发送的用于表明加入该组播组的报告消息，就刷新该定时器，如果该定时器溢出，路由器将在接口记录的组播组列表中删除该组播组。定时器的溢出时间大于 2×普遍查询消息发送间隔。

3. 离开组播组

如果某个终端离开了原先加入的某个组播组，向路由器发送离开消息，离开消息中给出标识该组播组的组播地址，并以该组播地址为封装离开消息的 IP 分组的目的 IP 地址。路由器接收到该离开消息后，发送指定组播组查询消息，指定的组播组就是该终端离开的组播组。如果网络中还存在加入了该组播组的终端，该终端将向路由器发送表明加入该组播组的报告消息，路由器接收到该报告消息后，只是刷新该组播组关联的定时器。如果路由器发送指定组播组查询消息后，规定时间内一直接收不到表明加入该组播组的报告消息，表明网络中不存在加入该组播组的终端，路由器将在接口记录的组播组列表中删除该组播组。

7.2.3 IGMP 侦听

在启动 IGMP 侦听功能前，交换机广播以 MAC 组地址为目的地址的 MAC 帧，即使网络中只有一个终端加入某个组播组，发送给属于该组播组的终端的组播 IP 分组也将在网络中广播，这将极大地浪费网络带宽，增加终端的处理负担。因此，交换机必须建立组播表，通过交换机建立的组播表将以 MAC 组地址为目的地址的 MAC 帧的传播范围控制在属于该 MAC 组地址对应的组播组的终端。交换机 IGMP 侦听功能是指交换机的这样一种功能，当交换机接收到以 MAC 组地址为目的地址的 MAC 帧时，首先判别该 MAC 帧的净荷是否是 IP 分组，在确定该 MAC 帧的净荷是 IP 分组的前提下，再判别该 IP 分组的净荷是否是 IGMP 消息，在确定该 IP 分组净荷是 IGMP 消息的前提下，对 IGMP 消息进行深入分析，并根据分析结果和接收该 IGMP 消息的端口构建组播表。

1. IGMP 消息封装过程

1）IP 组播地址映射到 MAC 组地址过程

以太网 MAC 地址分为三类：单播地址、组地址和全 1 表示的广播地址。单播地址类型和组播地址类型通过 MAC 地址高字节中的第 0 位进行区分，如果该位为 1，表明是组地址，如果该位为 0，表明是单播地址。IP 组播地址映射为 MAC 组地址的过程如图 7.8 所示。

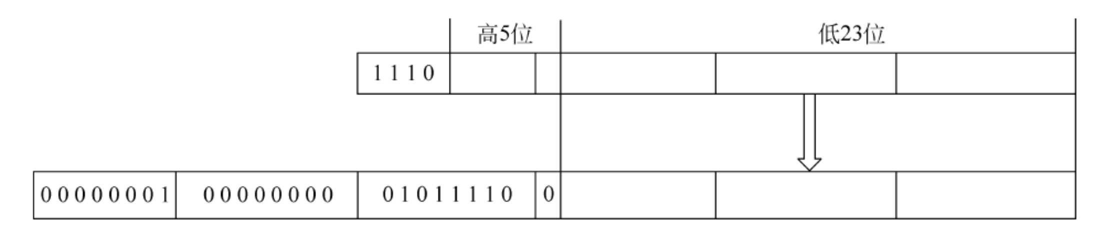


图 7.8 IP 组播地址映射到 MAC 组地址的过程

从图 7.8 中可以看出，映射后的 MAC 地址的高 25 位固定为 00000001、00000000、01011110 和 0，低 23 位等于组播地址的低 23 位。由于 IP 组播地址中用于标识组播组的地址有 28 位，因此，标识组播组的 IP 组播地址中的高 5 位在映射过程中没有使用，这就使得 IP 组播地址和 MAC 组地址之间的映射不是唯一的，32 个不同的 IP 组播地址有可能映射为同一个 MAC 组地址。图 7.9 给出了 32 个不同的 IP 组播地址（224.85.170.170、224.213.170.170、225.85.170.170、…、239.213.170.170）映射到同一个 MAC 组地址（01-00-5E-55-AA-AA）的过程。

高5位　低23位

1110	xxxx	x	1010101	10101010	10101010

00000001	00000000	01011110	0	1010101	10101010	10101010

图 7.9 32 个不同的 IP 组播地址映射为同一个 MAC 组地址

2）IGMP 消息封装成 MAC 帧过程

图 7.10 所示是将 IGMP 指定组播组查询消息封装成 MAC 帧过程，假定路由器发送 IGMP 消息的接口 IP 地址为 192.1.1.253，MAC 地址为 000C.85ED.8C01，标识指定组播组的 IP 组播地址为 237.37.37.37，则该 IGMP 消息首先被封装成以 192.1.1.253 为源 IP 地址、237.37.37.37 为目的 IP 地址的组播 IP 分组。将该组播 IP 分组封装成 MAC 帧时，首先需要将 IP 组播地址 237.37.37.37 映射到 MAC 组地址，根据图 7.8 所示的映射过程，IP 组播地址 237.37.37.37 映射到 MAC 组地址 0100.5E25.2525，因此，该 IP 组播分组被封装成源 MAC 地址为 000C.85ED.8C01、目的 MAC 地址为 0100.5E25.2525 的 MAC 帧，通过类型字段值 0800 表明 MAC 帧的净荷是 IP 分组。

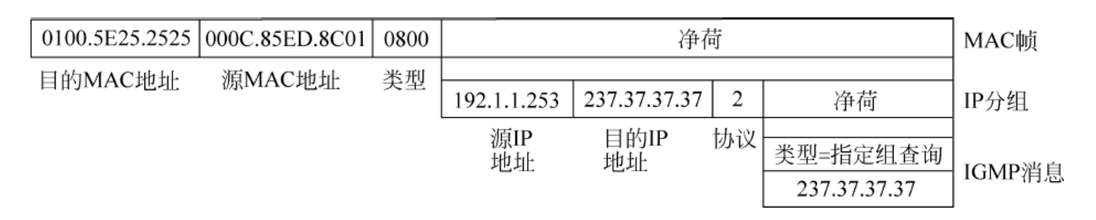


图 7.10　IGMP 消息封装成 MAC 帧过程

2. 确定连接路由器端口

交换机使能 IGMP 侦听功能后，如果接收到某个 IGMP 消息，且该 IGMP 消息的类型是普遍查询消息或指定组播组查询消息，将接收该 IGMP 消息的端口确定为连接路由器的端口，并将从除连接路由器端口以外的所有其他端口广播封装普遍查询消息而成的 MAC 帧。封装指定组播组查询消息的 MAC 帧根据交换机建立的组播表进行转发。

3. 创建组播转发项

如果交换机通过某个端口接收到报告消息，根据报告消息中给出的 IP 组播地址查找组播表，如果在组播表找不到该 IP 组播地址关联的组播转发项，创建组播转发项，组播地址为报告消息中给出的 IP 组播地址（或是该 IP 组播地址对应的 MAC 组地址），输出端口列表为连接路由器端口和接收该报告消息端口。向路由器转发该报告消息，启动定时器，在定时器溢出前，不再向路由器转发表明加入相同组播组的报告消息。需要强调的是，交换机一旦从接收到的以 MAC 组地址为目的地址的 MAC 帧中分离出 IGMP 消息，立即将该 IGMP 消息提交给交换机 IGMP 实体，由交换机 IGMP 实体根据 IGMP 消息类型分别进行处理，不会简单广播封装了 IGMP 消息的 MAC 帧，因此，直接连接在交换机上的终端是接收不到其他终端发送的报告消息的，为了避免每一个终端单独向路由器发送报告消息，交换机通过上述机制保证只把第一个终端发送的报告消息转发给路由器，其他终端发送的表明加入相同组播组的报告消息不再转发给路由器。

4. 添加输出端口

如果交换机通过某个端口接收到报告消息，并在组播表中找到与报告消息中给出的 IP 组播地址（或是该 IP 组播地址对应的 MAC 组地址）关联的组播转发项，如果组播转发项的

输出端口列表中没有包含接收该报告消息的端口,将接收该报告消息的端口添加到输出端口列表中,如果已经为表明加入相同组播组的报告消息启动了定时器,交换机不再向路由器转发该报告消息,否则,向路由器转发该报告消息,并启动定时器。

5. 普通离开组播组

如果交换机通过某个端口接收到离开消息,交换机通过接收离开消息的端口发送IGMP指定组播组查询消息,指定的组播组为终端表明离开的组播组。如果在规定时间内一直没有通过该端口接收到表明加入该组播组的报告消息,交换机将在该组播组对应的组播转发项的输出端口列表中删除接收到离开消息的端口。如果删除接收到离开消息的端口后,该组播组对应的组播转发项的输出端口列表中只剩下连接路由器端口,交换机将向路由器转发该离开消息,并从组播表中删除该组播转发项。

6. 立即离开组播组

如果某个交换机端口直接连接终端,如图7.7中的交换机S1和交换机S2,通过配置可以使交换机工作在立即离开方式,一旦某个终端离开某个组播组,该终端向交换机发送离开消息,离开消息中给出终端离开的组播组。交换机接收到离开消息后,立即从该组播组关联的组播转发项的输出端口列表中删除连接该终端的交换机端口。和普通离开组播组过程相同,如果删除接收到离开消息的端口后,该组播组对应的组播转发项的输出端口列表中只剩下连接路由器端口,交换机将向路由器转发该离开消息,并从组播表中删除该组播转发项。

7. 以太网组播IP分组传输过程

如果路由器R1和路由器R2接口1的IP地址配置如图7.11所示,路由器R1成为查询者,通过接口1周期性发送普遍查询消息,根据各个终端响应的报告消息记录如表7.4所示的组播组情况。交换机通往路由器R1接口1的端口成为连接路由器端口,各个交换机生成的组播表如图7.11所示。当路由器R1通过接口1发送属于组播组G1的组播IP分组时,该组播IP分组的传输过程如图7.11所示。虽然封装属于组播组G1的组播IP分组

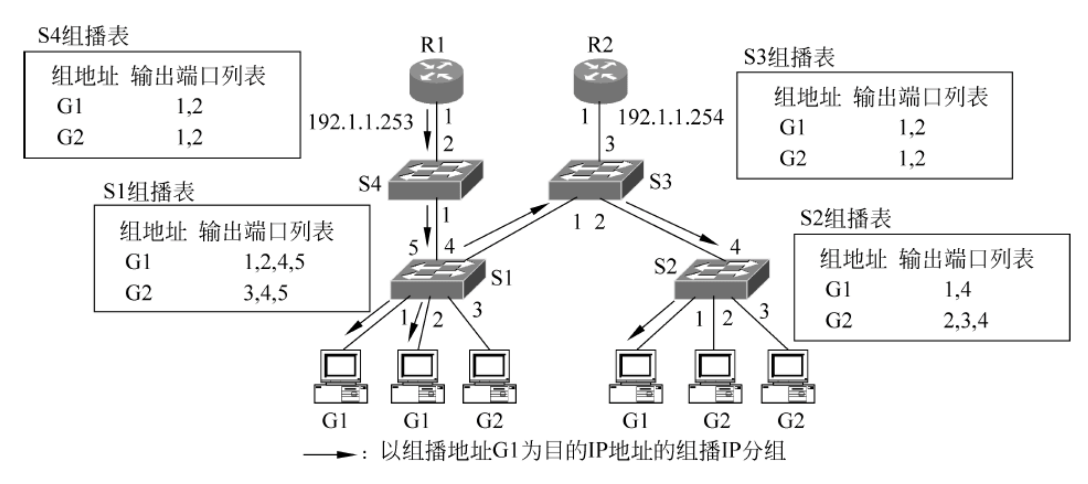


图7.11 交换机组播表

的 MAC 帧的目的 MAC 地址和封装用于表明加入组播组 G1 的 IGMP 报告消息的 MAC 帧的目的 MAC 地址相同，但交换机通过深入分析组播 IP 分组，确定组播 IP 分组净荷是 IGMP 消息或是普通数据，根据建立的组播表转发封装普通数据的组播 IP 分组，从封装 IGMP 消息的组播 IP 分组中提取出 IGMP 消息，并将 IGMP 消息提交给 IGMP 实体进行处理。只是由于交换机需要深入分析组播 IP 分组的净荷后，才能确定是封装普通数据的组播 IP 分组，还是封装 IGMP 消息的组播 IP 分组，使得 IGMP 侦听对交换机性能的影响非常大，使能 IGMP 侦听，会影响交换机转发 MAC 帧的速率。

表 7.4 路由器 R1 记录的组播组情况

路由器接口	接口连接的网络中终端加入组播组的情况
路由器 R1 接口 1	G1、G2

7.3 组播路由协议

路由器为了实现组播，必须建立组播路由表，组播路由表中的每一项组播路由项由三部分组成：一是源终端地址和组播组对(S,G)，二是上游接口，三是下游接口列表。当路由器接收到某个组播 IP 分组，首先在组播路由表中检索组播路由项，如果某项组播路由项中的 S 与该组播 IP 分组的源 IP 地址最长匹配，且该组播路由项中的 G 与组播 IP 分组的目的 IP 地址精确匹配，表示该组播路由项与该组播 IP 分组匹配，只有当路由器通过匹配的组播路由项中的上游接口接收该组播 IP 分组时，通过下游接口列表中给出的所有接口输出该组播 IP 分组。因此，路由器实现组播的关键是建立组播路由表，组播路由协议就是用于路由器在通过 IGMP 了解属于各个组播组的终端的分布情况的基础上，构建用于指明通往连接属于各个组播组的终端的网络的最短路径的组播路由项。

7.3.1 DVMRP

距离向量组播路由协议(Distance Vector Multicast Routing Protocol，DVMRP)是一种用于在路由器中构建组播路由表的组播路由协议。对于特定的源终端 S，首先建立源终端 S 通往所有网络的最短路径，然后，各个路由器通过 IGMP 了解直接连接的网络中的终端加入组播组的情况，通过剪枝过程，每一个路由器在已经建立的特定源终端 S 至所有网络的最短路径的基础上，建立特定源终端 S 至所有连接属于特定组播组 G 的终端的网络的最短路径，完成(S,G)对关联的组播路由项的创建过程。

1. DVMRP 消息类型和格式

1) DVMRP 消息类型及作用

DVMRP 包括探测消息、报告消息、剪枝消息、嫁接消息、嫁接应答消息等，探测消息用于发现邻居、监测邻居状态、同步路由器刚启动时和邻居的组播路由项。报告消息用于向邻居报告路由器组播路由表中的全部组播路由项，并传输毒性反转信息，相邻路由器之间通过周期性地交换报告消息完成组播路由表的创建。剪枝消息用于在以源终端(或源网络)S 为

根的广播树上，针对特定的组播组 G 剪去不需要传输以标识该组播组的组播地址为目的地址的组播 IP 分组的分枝，以此构建(S,G)对关联的组播树。嫁接消息的作用刚好相反，将通过剪枝消息剪去的分枝重新嫁接到(S,G)对关联的组播树上，嫁接应答消息用于对接收到的嫁接消息做出确认应答。

2）DVMRP 消息格式

图 7.12 给出了报告消息和剪枝消息格式，DVMRP 消息被封装成组播 IP 分组，源 IP 地址是发送该 DVMRP 消息的路由器接口的 IP 地址，目的 IP 地址是组播地址 224.0.0.4，表明接收端是接口所连接网络中所有运行 DVMRP 的路由器。协议字段值 2 表明 IP 分组净荷是 IGMP 消息，因此，所有 GVMRP 消息成为 IGMP 消息的一种类型，图中用“类型＝DVMRP 消息”表示，通过编码来区分不同的 DVMRP 消息。报告消息给出组播路由项，每一项组播路由项给出该路由器到达特定源终端(或源网络)的距离。报告消息由于包含路由器的全部组播路由项，因而被称为路由消息。路由器发送 DVMRP 报告消息的时机与路由器发送 RIP 路由消息的时机基本相同。

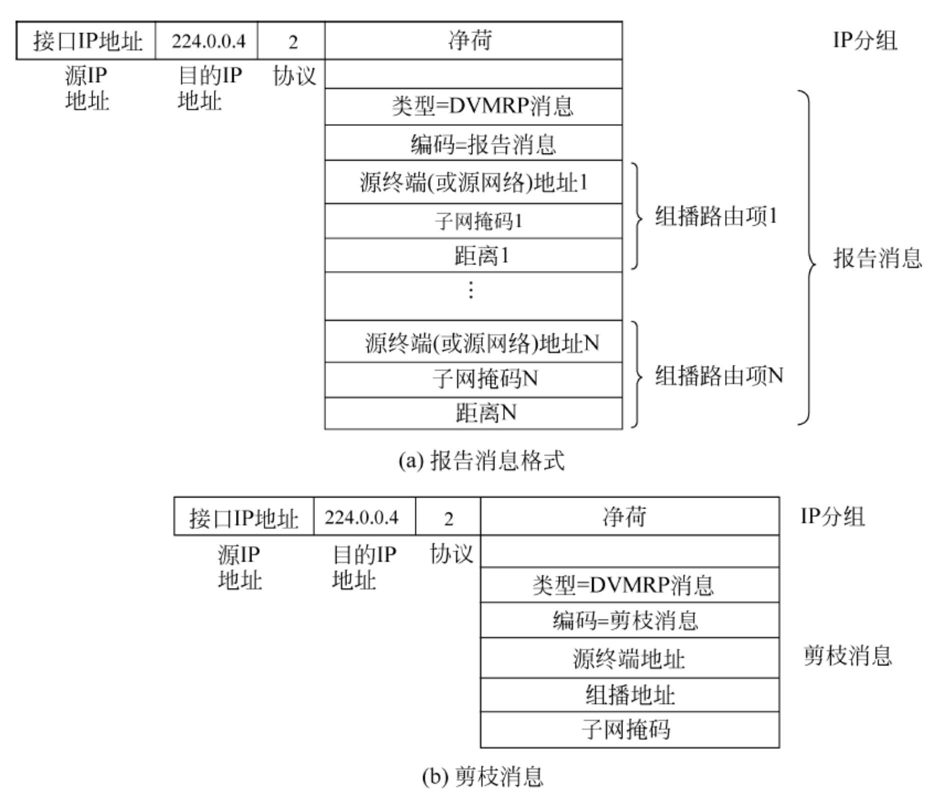


图 7.12　部分 GVMRP 消息格式

剪枝消息用于在以源终端(或源网络)S 为根的广播树上，针对特定的组播组 G 剪去不需要传输以标识该组播组的组播地址为目的 IP 地址的组播 IP 分组的分枝，通常由连接末端网络的叶路由器在接收到源终端发送的以组播地址 G 为目的地址的组播 IP 分组后，且发现直接连接的网络中不存在属于组播组 G 的终端时，通过发送剪枝消息开始剪枝过程。剪枝消息中的源终端地址是接收到的以组播地址 G 为目的地址的组播 IP 分组的源 IP 地址，如果该路由器可以得出源终端所在网络(称为源网络)的子网掩码，作为可选项可以给出

源网络的子网掩码，组播地址指定某个组播组，表示不需要向该分枝传输以该组播地址为目的 IP 地址的组播 IP 分组。叶路由器始发的剪枝消息沿着通往源终端(或源网络)的传输路径逐跳转发，路由器接收到剪枝消息后，用剪枝消息给出的源终端(或源网络)地址 S 和组播地址 G 检索组播路由表，找到(S,G)匹配的组播路由项，只有确定该剪枝消息通过该组播路由项下游接口列表中的其中一个接口接收时，才对该剪枝消息进行处理，否则，丢弃该剪枝消息。

2. 广播树建立过程

1) 基本思路

以源终端 S 为根，源终端 S 通往所有网络的最短路径为分枝的树，就是源终端 S 对应的广播树。对应图 7.3 所示的互连网络结构中，源终端 S 至其他三个末端网络的广播树如图 7.4 所示。如果某个路由器 I 与路由器 J 和路由器 K 相邻，且路由器 J 在源终端 S 至某个网络的最短路径中先于路由器 I，路由器 K 在源终端 S 至某个网络的最短路径中后于路由器 I，称路由器 J 为路由器 I 的前一跳路由器(也称上游路由器)，路由器 K 为路由器 I 的下一跳路由器(也称下游路由器)，为了保证组播 IP 分组沿着源终端 S 对应的广播树传播，路由器只转发从连接前一跳路由器的接口进入的组播 IP 分组，而且，只通过连接下一跳路由器的接口输出该组播 IP 分组。路由器 R5 针对图 7.4 所示的广播树，只转发从接口 1 进入的组播 IP 分组，且只从接口 2 和接口 3 输出该组播 IP 分组。因此，针对特定源终端 S 建立广播树的过程就是在每一个路由器建立确定连接源终端 S 至所有网络最短路径上的前一跳和下一跳路由器的接口的过程。某个路由器连接源终端 S 至所有网络最短路径上的前一跳路由器的接口，称为上游接口，连接下一跳路由器的接口称为下游接口，不同源终端对应的源终端至所有网络的最短路径是不同的，图 7.13 给出 IP 地址为 192.1.2.1 的源终端至所有其他末端网络的最短路径，因此，每一个路由器需要在组播路由表中为不同的源终端建立一项组播路由项，该组播路由项中给出该源终端至所有网络最短路径对应的上游接口和下游接口列表。对于图 7.4 和图 7.13 所示的广播树，路由器 R5 需要建立表 7.5 所示的组播路由表。表中的源网络给出源终端所在的网络。在讨论 DVMRP 建立的组播路由项时，假定直接连接的网络的距离为 1。

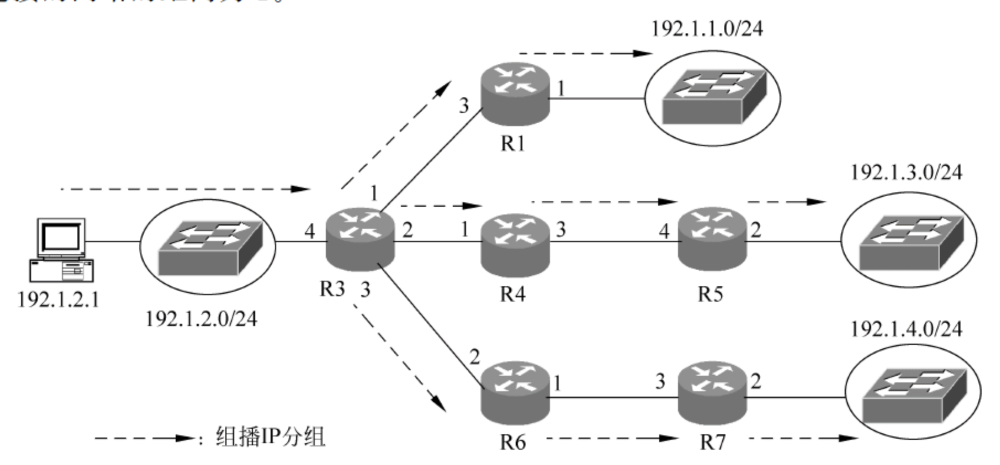


图 7.13 以网络 192.1.2.0/24 中终端为源终端的广播树

表 7.5 路由器 R5 组播路由表

源网络	距离	前一跳路由器	上游接口	下游接口列表
192.1.1.0/24	3	193.1.4.1	1	2,3
192.1.2.0/24	3	193.1.7.1	4	2

每一个路由器建立类似表 7.5 所示的组播路由项的思路如下：如果传输路径是对称的，即源终端 S 至该路由器的最短路径和该路由器至源终端 S 的最短路径相同，对于路由器 R 而言，单播路由表中以源终端 S 所在网络为目的网络的路由项中的下一跳路由器就是源终端 S 至路由器 R 的最短路径上的前一跳路由器，如果源终端 S 至某个网络的最短路径经过路由器 R，则源终端 S 至路由器 R 的最短路径上的前一跳路由器就是源终端 S 至该网络的最短路径上的前一跳路由器。因此，只要某个路由器位于源终端 S 至某个网络的最短路径上，根据该路由器的单播路由表可以确定源终端 S 至该网络的最短路径上的前一跳路由器和上游接口。确定路由器下游接口的工作也比较简单，如果该路由器直接连接某个末端网络，则连接该末端网络的接口就是源终端 S 对应的下游接口。对于连接其他类型网络的接口，通过下述方法确定该接口是否是源终端 S 对应的下游接口。如果某个路由器 R 通过单播路由表确定源终端 S 至某个网络最短路径上的前一跳路由器 X，则路由器 X 连接路由器 R 的接口就是路由器 X 的下游接口，路由器 R 可以通过连接路由器 X 的接口发送一个特定路由项，该路由项的源终端为 S，距离为特定值，当路由器 X 通过某个接口接收到源终端为 S 的特定路由项，确定该接口为源终端 S 对应的下游接口。

2）DVMRP 建立组播路由表过程

DVMRP 的工作思路和 RIP 相似，都是找出某个路由器通往特定网络的最短路径，只是最短路径的含义不同，在组播路由表中，最短路径表示属于该网络的源终端至该路由器的最短路径，这样，表 7.6 所示的路由器 R5 的单播路由表可以直接转换成表 7.7 所示的路由器 R5 的组播路由表。

表 7.6 路由器 R5 单播路由表

目的网络	距离	下一跳路由器	输出接口
192.1.1.0/24	3	193.1.4.1	1
192.1.2.0/24	3	193.1.7.1	4
192.1.3.0/24	1	直接	2
192.1.4.0/24	2	193.1.9.2	3

表 7.7 路由器 R5 组播路由表

源网络	距离	前一跳路由器	上游接口	下游接口列表
192.1.1.0/24	3	193.1.4.1	1	
192.1.2.0/24	3	193.1.7.1	4	
192.1.3.0/24	1	—	2	
192.1.4.0/24	2	193.1.9.2	3	

DVMRP 求出表 7.7 中通往源终端所在网络的最短路径、前一跳路由器及上游接口的过程，和 RIP 求出表 7.6 所示的通往目的网络的最短路径、下一跳路由器和输出接口的过

程完全相同，值得强调的是 DVMRP 求出表 7.7 中下游接口列表一栏中的内容的过程。下游接口分为两类：一类是直接连接末端网络的接口；另一类接口是作为源终端至某个路由器的最短路径上的前一跳路由器用于连接该路由器的接口。如图 7.4 所示的广播树中，路由器 R5 作为源终端至路由器 R7 的最短路径上的前一跳路由器，而接口 3 是路由器 R5 用于连接路由器 R7 的接口。如果对于特定的源终端，某个路由器的下游接口都是直接连接末端网络的接口，该路由器被称为是以该源终端为根的广播树的叶路由器。由于 DVMRP 是类似 RIP 的距离向量路由协议，因此，路由器 R5 本身是无法通过 DVMRP 推导出下游接口的，但路由器 R7 可以通过 DVMRP 得出源终端至路由器 R7 的最短路径上的前一跳路由器——R5，并可以通过毒性反转技术通知路由器 R5：它是源终端至路由器 R7 的最短路径上的前一跳路由器，路由器 R5 将接收到毒性反转信息的接口作为下游接口。毒性反转技术是指某个路由器如果从接口 X 接收到的路由消息（DVMRP 报告消息）中推导出组播路由项 Y，可以断定接口 X 是组播路由项 Y 的上游接口，组播路由项 Y 对应的前一跳路由器和接口 X 相连，因此，通过接口 X 发送一个包含组播路由项 Y 的路由消息，但将组播路由项 Y 的距离值设置成一个特殊值。当前一跳路由器接收到包含特殊距离值的组播路由项 Y 后，就将接收该路由消息的接口作为组播路由项 Y 对应的下游接口。

DVMRP 为了求出每一个组播路由项对应的下游接口列表，要求在发送给某个相邻路由器的路由消息中包含从该相邻路由器学习到的组播路由项，但用特殊的距离值标明这些从该相邻路由器学习到的组播路由项。DVMRP 以 32 作为无穷大值，用“32＋距离值”表明该路由项是从发送该路由消息的接口接收到的路由消息中学习到的组播路由项。由于路由器 R4 的组播路由表中源网络为 192.1.3.0/24 和 192.1.4.0/24 的组播路由项是从通过接

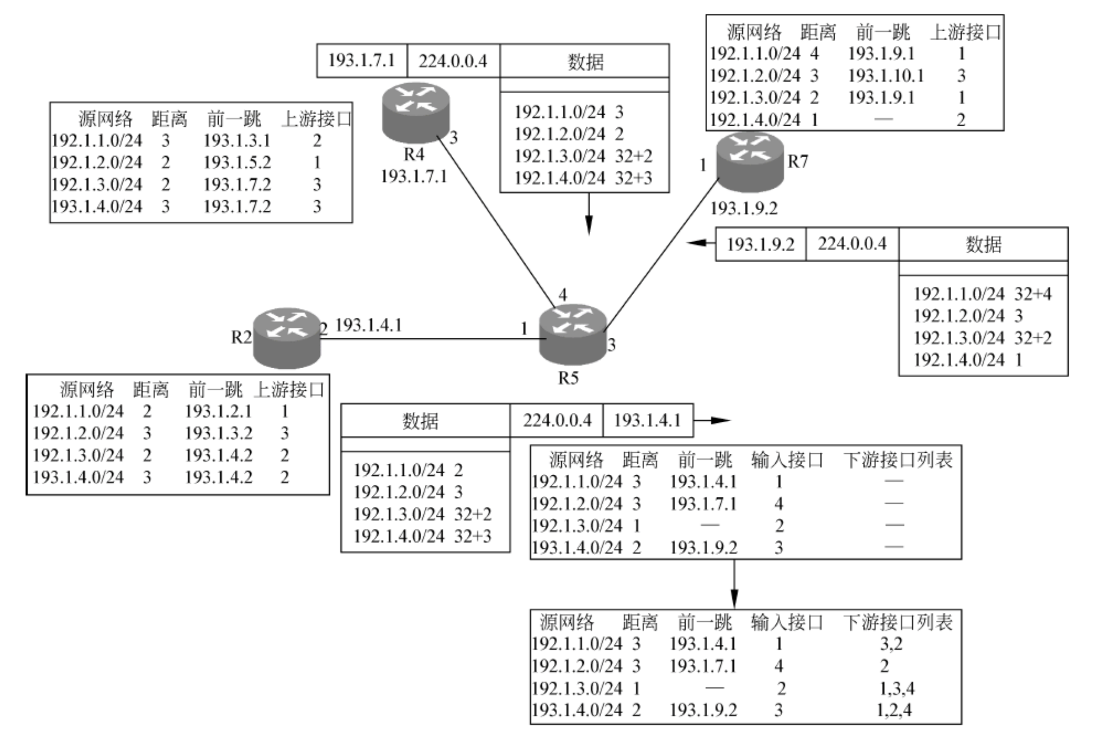


图 7.14　路由器 R5 生成最终组播路由表的过程

口 3 接收到的路由消息中学习到的,因此,路由器 R4 从接口 3 发送的路由消息中包含源网络为 192.1.3.0/24 和 192.1.4.0/24 的组播路由项,但距离值设置成 32+3(3 是该组播路由项中的距离)。当路由器 R5 通过接口 4 接收到该路由消息后,发现源网络为 192.1.3.0/24 和 192.1.4.0/24 的组播路由项的距离值设置成 32+3,确定路由器 R4 是通过自己发送的路由消息学习到这两项组播路由项,自己是组播路由项指定的源网络至路由器 R4 的最短路径上的前一跳路由器,将接收该路由消息的接口(接口 4)设置成这两项组播路由项的下游接口,如图 7.14 所示。当和路由器 R5 相邻的路由器均通过毒性反转技术将路由器 R5 作为前一跳路由器的组播路由项告知路由器 R5,路由器 R5 建立图 7.14 所示的所有组播路由项的下游接口列表内容。最终生成的组播路由表如表 7.8 所示。

表 7.8 路由器 R5 组播路由表

源网络	距离	前一跳路由器	上游接口	下游接口列表
192.1.1.0/24	3	193.1.4.1	1	2,3
192.1.2.0/24	3	193.1.7.1	4	2
192.1.3.0/24	1	—	2	1,3,4
192.1.4.0/24	2	193.1.9.2	3	1,2,4

图 7.15 所示的网络结构中,路由器 R1、路由器 R2 和路由器 R3 连接在同一个以太网上,对于这样的连接方式,三个路由器中只能有一个路由器转发来自源终端 S 的组播 IP 分组,该路由器称为该以太网中针对源终端 S 的指定路由器,确定指定路由器的原则如下:

- 所有路由器中到达源终端 S 的距离最小的路由器;
- 如果存在多个到达源终端 S 的最小距离相同的路由器,其中连接该以太网接口的 IP 地址最小的路由器。

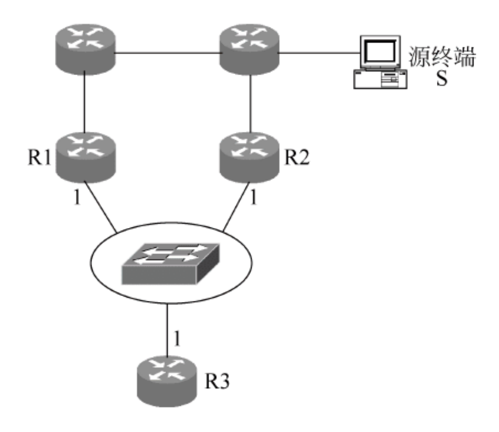


图 7.15 指定路由器含义

由于每一个路由器通过连接该以太网接口发送的路由消息(DVMRP 报告消息)能被连接在同一以太网的所有其他路由器接收,因此,每一个路由器通过接收其他路由器发送的路由消息不难确定自己是否是该以太网中针对源终端 S 的指定路由器。图 7.15 中,由于路由器 R2 到达源终端 S 的距离最短,路由器 R2 成为该以太网中针对源终端 S 的指定路由器。

如图 7.15 所示,路由器 R3 根据通过接口 1 接收的路由消息构建源终端 S 对应的组播路由项,因此,在通过接口 1 发送的路由消息中,源终端 S 对应的组播路由项的距离是 32+X,由于路由器 R3 接口 1 连接的是以太网,因此,路由器 R1 和路由器 R2 的接口 1 均接收到路由器 R3 通过接口 1 发送的路由消息,但只有作为该以太网中针对源终端 S 的指定路由器的路由器 R2 能够把接口 1 作为源终端 S 对应的组播路由项的下游接口。

3. 剪枝

DVMRP 建立广播树对应的组播路由表时,并不考虑属于各个组播组的终端的分布情

况，因此，对于图 7.16 所示的属于各个组播组的终端的分布情况，DVMRP 首先建立如图 7.16 所示的指定源终端至所有末端网络的最短路径的广播树。

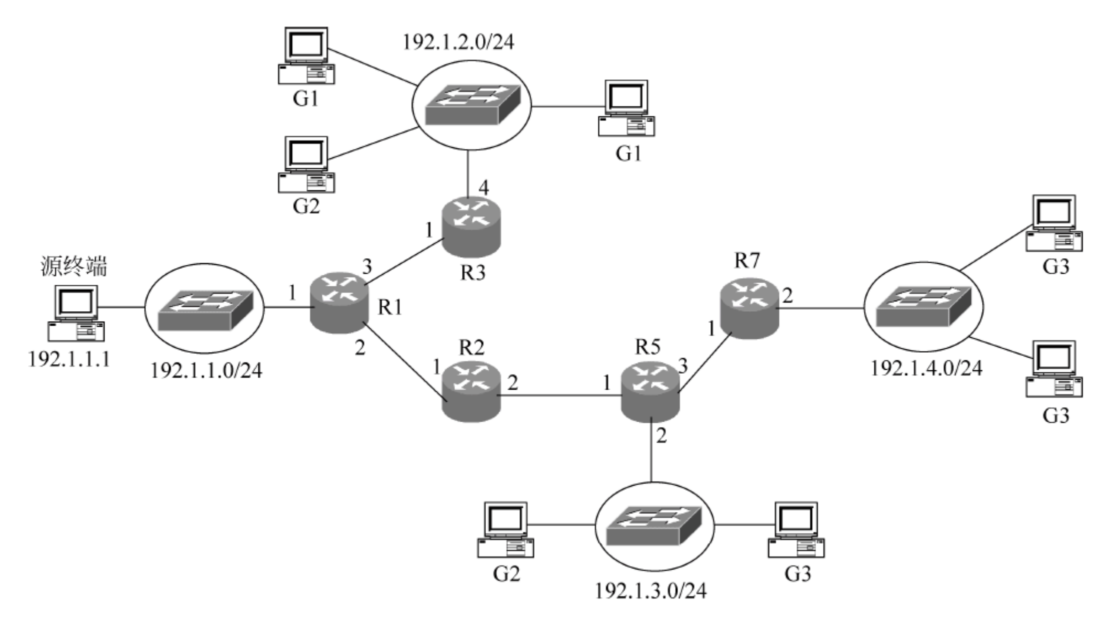


图 7.16　多个组播组并存的情况

各个路由器通过 IGMP 了解直接连接的网络中的终端加入组播组的情况，对应图 7.16 所示的属于各个组播组的终端的分布情况，直接连接末端网络的路由器 R3、路由器 R5 和路由器 R7 记录如表 7.9 所示的直接连接的末端网络中终端加入组播组的情况。

表 7.9　连接末端网络的接口记录的信息

路由器接口	直接连接的网络中的终端加入组播组的情况
路由器 R3 接口 4	G1，G2
路由器 R5 接口 2	G2，G3
路由器 R7 接口 2	G3

用 DVMRP 建立的属于网络 192.1.1.0/24 的源终端对应的广播树如图 7.16 所示，源终端发送的分别以组播地址 G1、组播地址 G2 和组播地址 G3 为目的 IP 地址的组播 IP 分组将到达图 7.16 中的所有路由器，这将极大地浪费链路带宽。对于以源终端（或源网络）S 为根的广播树，如果某个接口连接的分枝中没有存在属于组播组 G 的终端，称该接口与（S，G）不匹配，一旦该接口和（S，G）不匹配，将截断源终端 S 发送的、目的地址为组播地址 G 的组播 IP 分组，这个过程称为针对（S，G）的剪枝过程，即在以源终端（或源网络）S 为根的广播树中剪除不需要传输以组播地址 G 为目的地址的组播 IP 分组的分枝。针对图 7.16 所示广播树的剪枝过程如下：属于源网络 192.1.1.0/24 的源终端发送的、以组播地址 G1、组播地址 G2 和组播地址 G3 为目的 IP 地址的第一个组播 IP 分组到达图 7.16 中的所有路由器，当路由器 R3 接收到以组播地址 G3 为目的 IP 地址的组播 IP 分组时，发现接口 4 连接的网络中没有加入组播组 G3 的终端，而且路由器 R3 又是叶路由器，即广播树中该分枝的最后一个路由器。路由器 R3 向它的前一跳路由器发送剪枝消息，剪枝消息中给出源终端地址 192.1.1.1 和组播地址 G3，表明该分枝不需转发以组播地址 G3 为目的 IP 地址的组播

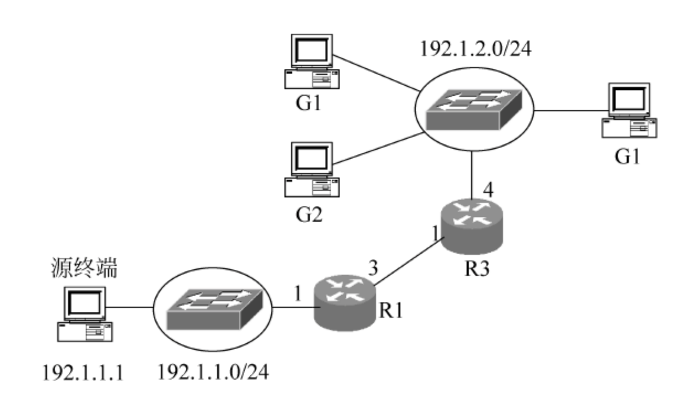
IP 分组。路由器 R1 通过接口 3 接收到该剪枝消息，用剪枝消息给出的源终端地址 192.1.1.1检索组播路由表，匹配源网络为 192.1.1.0/24 的组播路由项，确定接口 3 在该组播路由项的下游接口列表中，路由器 R1 将在接口 3 截断以组播地址 G3 为目的 IP 地址的组播 IP 分组。完成上述操作后，对于以组播地址 G3 为目的 IP 地址的组播 IP 分组，图 7.16 所示广播树中和路由器 R1 的接口 3 连接的分枝已被剪除。同样，路由器 R7 向前一跳路由器 R5 发送表明不需转发以组播地址 G1、组播地址 G2 为目的 IP 地址的组播 IP 分组的剪枝消息，路由器 R5 将在接口 3 截断以组播地址 G1、组播地址 G2 为目的 IP 地址的组播 IP 分组。由于路由器 R5 接口 2 直接连接的末端网络中也没有属于组播地址 G1 的终端，路由器 R5 向它的前一跳路由器发送表明不需转发以组播地址 G1 为目的 IP 地址的组播 IP 分组的剪枝消息。路由器 R2 将在接口 2 截断以组播地址 G1 为目的 IP 地址的组播 IP 分组，由于路由器 R2 连接的所有分枝均要求截断以组播地址 G1 为目的 IP 地址的组播 IP 分组，路由器 R2 向它的前一跳路由器发送表明不需转发以组播地址 G1 为目的 IP 地址的组播 IP 分组的剪枝消息，使路由器 R1 在接口 2 截断以 G1 为目的 IP 地址的组播 IP 分组。经过这一轮剪枝消息的传输，对图 7.16 所示的以源网络 192.1.1.0/24 为根的广播树，完成针对组播组 G1、组播组 G2 和组播组 G3 进行的剪枝过程，分别生成如图 7.17(a)、(b)和(c)所示的针对组播组 G1、组播组 G2 和组播组 G3 的组播树。之所以将剪枝后的广播树称为组播树是因为组播树只将以组播地址 G 为目的 IP 地址的组播 IP 分组传输给存在属于组播组 G 的终端的网络。

剪枝过程是针对以源终端(或源网络)S 为根的广播树展开的，在各个路由器直接连接的末端网络中的终端加入组播组的情况不变的前提下，不同源终端(或源网络)为根的广播树完成剪枝过程后的结果是不同的，因此，组播树是(S，G)相关的，即(S，G)组播树是在以源终端(或源网络)S 为根的广播树的基础上，针对组播组 G 完成剪枝过程后的结果。在图 7.16 所示的以源网络 192.1.1.0/24 为根的广播树上，针对组播组 G1、组播组 G2 和组播组 G3 完成剪枝过程后，路由器 R5 的组播路由表中和源网络 192.1.1.0/24 对应的组播路由项变成表 7.10 所示内容。

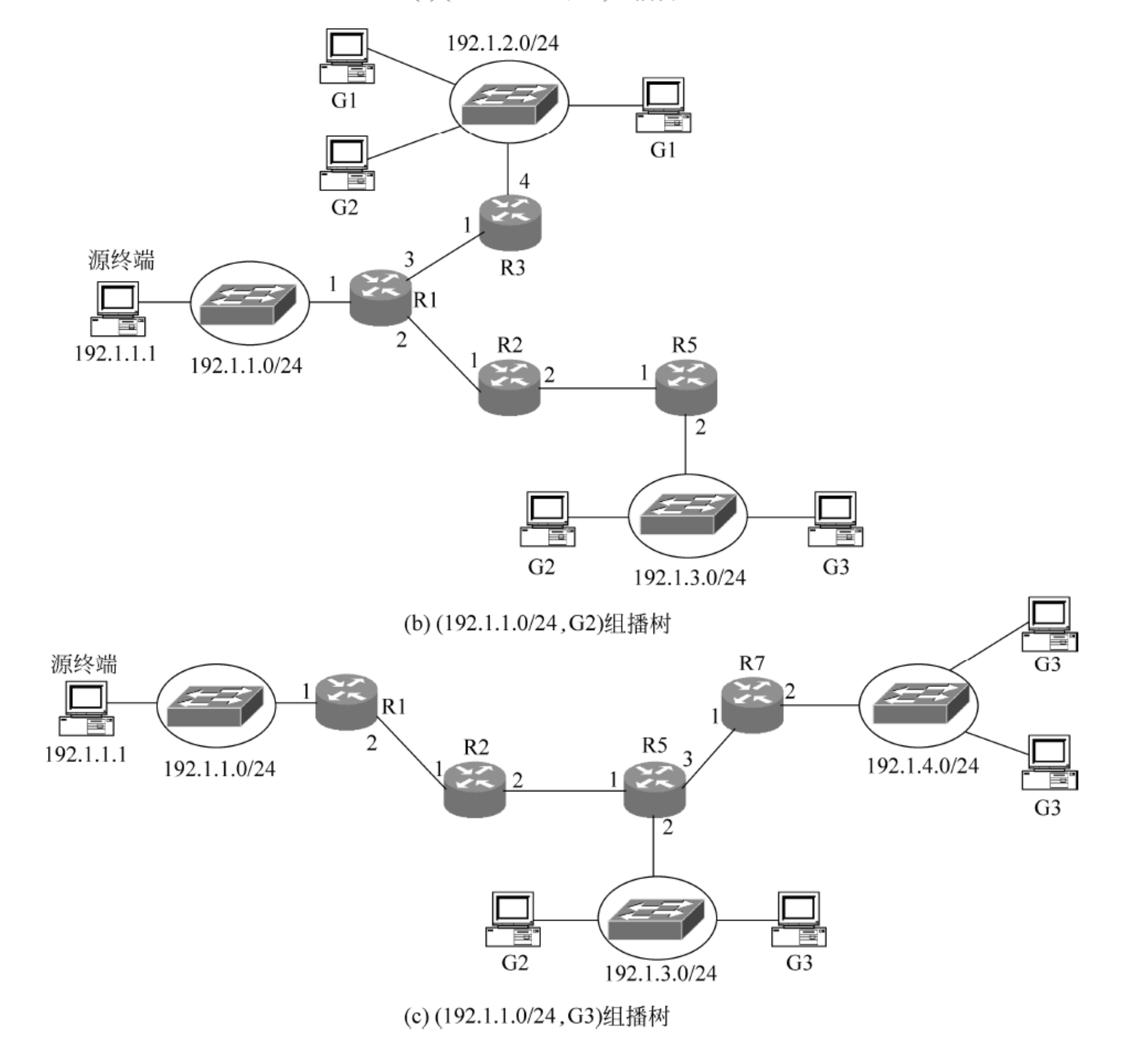
表 7.10 路由器 R5 对应特定组播组的组播路由表

源网络	组播组	前一跳路由器	距离	上游接口	下游接口列表
192.1.1.0/24	G1	193.1.4.1	3	1p	2p,3p
192.1.1.0/24	G2	193.1.4.1	3	1	2,3p
192.1.1.0/24	G3	193.1.4.1	3	1	2,3
192.1.2.0/24		193.1.7.1	3	4	2
192.1.3.0/24		—	1	2	1,3,4
192.1.4.0/24		193.1.9.2	2	3	1,2,4

对于表 7.10 中组播组 G1 对应的路由项。上游接口 1 后面的字符 p 意味着通过该接口向前一跳路由器发送了用以表明不需转发以组播地址 G1 为目的 IP 地址的组播 IP 分组的剪枝消息。下游接口后面的字符 p 表明接口所连接的末端网络中没有属于组播组 G 的终端，或者接口接收到用以表明不需转发以组播地址 G1 为目的 IP 地址的组播 IP 分组的剪枝消息。某个路由器向前一跳路由器发送某个组播组对应的剪枝消息的前提是该组播组关联的



(a) (192.1.1.0/24,G1)组播树



(b) (192.1.1.0/24,G2)组播树

(c) (192.1.1.0/24,G3)组播树

图 7.17　和特定源终端及组播组关联的组播树

组播路由项中的所有下游接口都被截断。表 7.10 中和组播组 G2 关联的组播路由项表明接口 3 接收到用以表明不需转发以组播地址 G2 为目的 IP 地址的组播 IP 分组的剪枝消息。

每一个下游接口的截断状态都是受定时器控制的，一旦在规定时间内接收不到对应的剪枝消息，将去除下游接口的截断状态。因此，组播路由项中标识 p 的上游接口必须周期性

地向前一跳路由器发送剪枝消息，当然，如果该路由项中的某个下游接口的状态从截断变为正常转发，上游接口将立即停止向前一跳路由器发送剪枝消息。由于前一跳路由器的下游接口从不再接收到剪枝消息到变为正常转发状态有一段时延，为了加快前一跳路由器下游接口从截断状态变为正常转发状态，可以向前一跳路由器发送嫁接消息，嫁接消息的作用和剪枝消息的作用刚好相反，如果某个路由器的下游接口接收到嫁接消息，该下游接口和特定组播组对应的状态立即从截断变为正常转发。

需要强调的是，DVMRP 生成的针对不同源终端（或源网络）的组播路由项只是构建了以这些源终端（或源网络）为根的广播树，真正生成以这些源终端（或源网络）为根的组播树，必须根据路由器直接连接的末端网络中终端加入组播组的情况，分别在以这些源终端（或源网络）为根的广播树上针对每一个互连网络中活跃的组播组完成剪枝过程。互连网络中活跃的组播组是指至少包含一个连接在互连网络上的终端的组播组。

4. 组播 IP 分组的传输过程

1）路由器转发组播 IP 分组的过程

源终端如果发送组播 IP 分组，构建以源终端 IP 地址为源 IP 地址，标识某个组播组的组播 IP 地址为目的 IP 地址的组播 IP 分组，该组播 IP 分组在源终端直接连接的网络中组播，到达连接源终端所在网络的路由器，该路由器接收到该组播 IP 分组后，在组播路由表中检索源网络和该组播 IP 分组的源 IP 地址最长匹配、组播地址和该组播 IP 分组的目的 IP 地址相等的组播路由项，在确定该组播 IP 分组通过匹配的组播路由项中的上游接口输入后，通过该组播路由项中所有接口状态为正常转发的下游接口输出该组播 IP 分组。

图 7.17(b)中，如果 IP 地址为 192.1.1.1 的源终端向所有属于组播组 G2 的终端发送组播 IP 分组，该组播 IP 分组的源 IP 地址为 192.1.1.1，目的 IP 地址为组播地址 G2。当路由器 R5 接收到该组播 IP 分组时，在表 7.10 所示的组播路由表中检索源网络和 192.1.1.1 匹配、组播地址等于 G2 的组播路由项，因为 IP 地址 192.1.1.1 属于网络地址 192.1.1.0/24，因此，检索结果是表 7.10 中和(192.1.1.0/24，G2)关联的组播路由项，在确定该组播 IP 分组通过该组播路由项中的上游接口（接口 1）输入后，将通过该组播路由项中接口状态为正常转发的所有下游接口输出该组播 IP 分组，由于该组播路由项中的下游接口中只有接口 2 的状态为正常转发，通过接口 2 输出该组播 IP 分组。

2）IP 组播地址对应的 MAC 组地址

假定路由器 R3 和终端之间的连接过程如图 7.18 所示，路由器 R3 如何通过以太网传输目的地址为组播地址 G1 的组播 IP 分组呢？路由器 R3 通过以太网传输组播 IP 分组前，须把组播 IP 分组封装成 MAC 帧，该 MAC 帧的源 MAC 地址是路由器 R3 接口 4 的 MAC 地址，目的 MAC 地址是组播地址 G1 对应的 MAC 组地址。

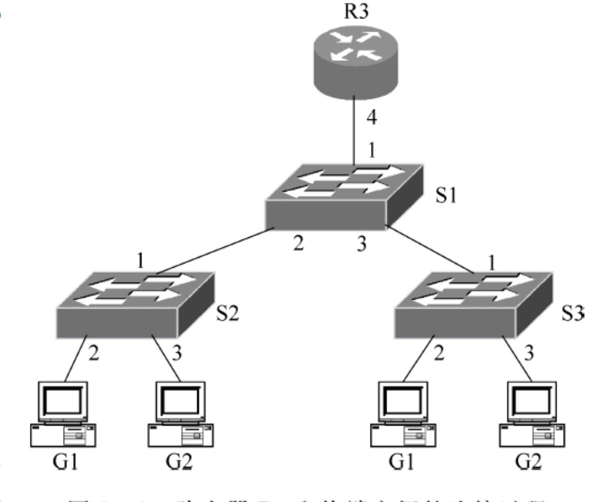


图 7.18　路由器 R3 和终端之间的连接过程

没有启动 IGMP 侦听功能前，以太网交换机对 MAC 组地址的处理过程等同于 MAC 广播地址(ff-ff-ff-ff-ff-ff)，如果 MAC 帧的目的 MAC 地址为 MAC 组地址，以太网交换机从除接收该 MAC 帧的端口以外的所有其他端口转发该 MAC 帧。因此，对于图 7.18 所示的网络结构，一旦路由器 R3 发送封装组播 IP 分组的 MAC 帧，该 MAC 帧以广播方式在以太网中传输。显然，这种传输方式无论对以太网带宽，还是终端都造成很大的压力。在讨论 IGMP 时已经讲到，当某个终端加入某个组播组时，向直接连接终端所在网络的路由器发送报告消息。当离开某个组播组时，向该路由器发送离开消息。封装报告或离开消息的 IP 分组是一个组播 IP 分组，目的 IP 地址就是标识终端要求加入或离开的组播组的组播地址。该组播 IP 分组被封装成 MAC 帧时，目的 MAC 地址就是该组播地址对应的 MAC 组地址。启动 IGMP 侦听功能后的以太网交换机对所有以这样的 MAC 组地址为目的地址的 MAC 帧进行分析，如果是报告消息，将输入该 MAC 帧的端口和该 MAC 帧的目的地址绑定在一起。如果是离开消息，则删除已经建立的绑定，整个过程如图 7.19 所示。以太网交换机通过 IGMP 侦听建立端口和组播组之间的绑定关系后，一旦接收到以 MAC 组地址为目的地址的 MAC 帧，以太网交换机先检索组播表，只从和该 MAC 组地址绑定的端口转发该 MAC 帧，如图 7.20 所示。

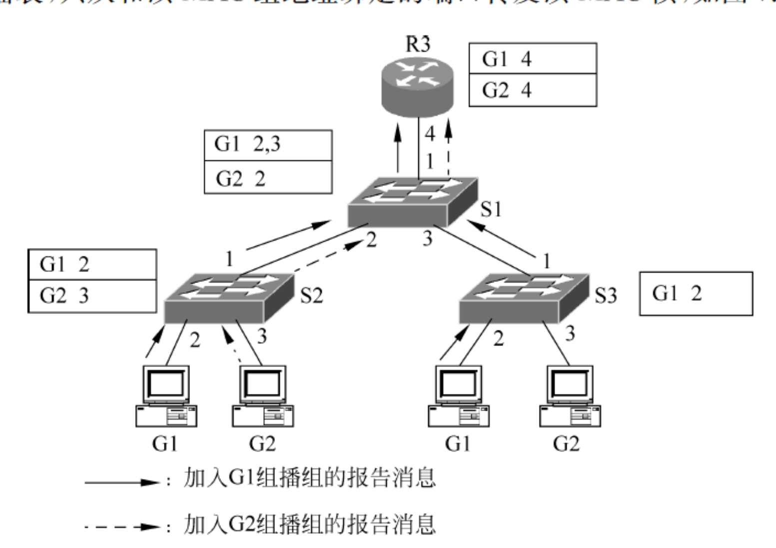


图 7.19　以太网交换机建立端口和 MAC 组地址之间绑定的过程

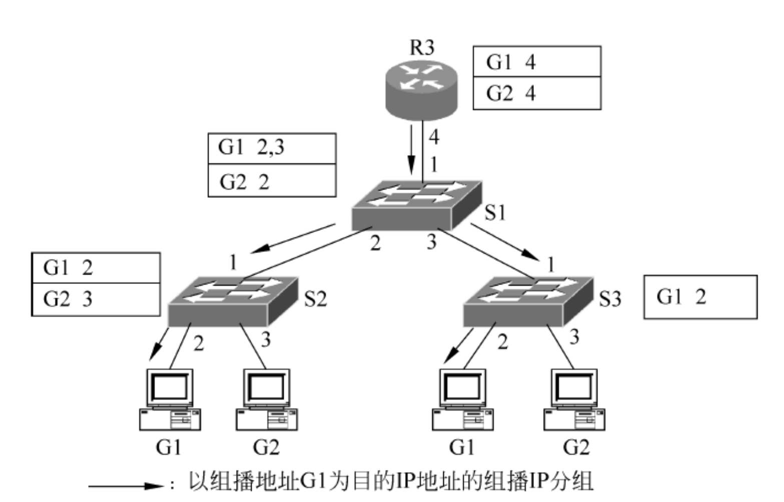


图 7.20　以太网组播以组地址为目的地址的 MAC 帧的过程

7.3.2 PIM-SM

DVMRP 是一种比较简单的组播路由协议，它不需要借助单播路由协议，如 RIP、OSPF 等内部网关协议，就可直接生成组播路由表。但从 7.3.1 节的讨论中可以看出：一是 DVMRP 是距离向量协议，它所得出的最短路径是最少跳数的传输路径，而用于确定链路代价的因素应该很多，不仅仅是经过的路由器跳数；二是 DVMRP 建立的是广播树，特定源终端发送的第一个组播 IP 分组将遍历互连网络中的所有路由器，在通过剪枝操作后，才将广播树剪枝成与特定组播组对应的组播树，如果在 Internet 中广播一个组播 IP 分组，其代价是无法想象的，因此，DVMRP 只能用于小规模网络。

目前有多种适用于不同组播应用环境的组播路由协议，如 PIM-SM（Protocol Independent Multicast-Sparse Mode，协议无关组播——稀疏方式）和 PIM-DM（Protocol Independent Multicast-Dense Mode，协议无关组播——密集方式）。这两种组播路由协议称为协议无关的组播路由协议是指源终端至其他终端的最短路径由其他单播路由协议建立，和组播路由协议无关，因此，最短路径的含义由对应的单播路由协议确定，和组播路由协议无关。这一点恰好弥补了 DVMRP 用最少跳数路径作为最短路径的缺陷。PIM-DM 适用环境和 DVMRP 相似，只能用于小规模互连网络且互连网络中的大部分终端都是组播组成员的组播应用环境。PIM-SM 与 DVMRP 和 PIM-DM 相反，适用于大规模互连网络且互连网络中只有少量终端是组播组成员的组播应用环境。互连网络结构和组播组成员分布如图 7.21 所示。

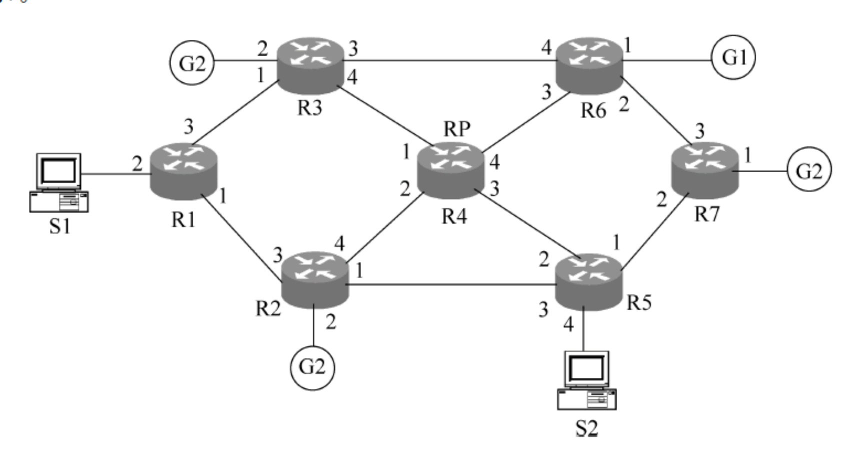


图 7.21 互连网络结构和组播组成员分布

1. 基本思路

PIM-SM 正常工作的前提是互连网络中的各个路由器已经建立单播路由表，但与建立单播路由表的路由协议无关。PIM-SM 在已经建立的单播路由表的基础上，为互连网络中活跃的组播组建立对应的组播树，所谓活跃的组播组是指至少包含了一个连接在互连网络上的终端的组播组，路由器通过 IGMP 掌握直接连接的网络中终端加入组播组的情况，因而掌握直接连接的网络中存在的活跃的组播组。但组播树是(S,G)相关的，如果针对互连网络中源终端（或源网络）和活跃的组播组的两两组合构建组播树，一是组播树的数量非常

庞大，二是由于大量源终端（或源网络）和活跃的组播组之间没有通信需求，会导致资源的极大浪费。因此，PIM-SM 为互连网络中活跃的组播组建立共享组播树，共享组播树的根是称为汇聚点（Rendezvous Point，RP）的路由器，某个源终端 S 如果需要向属于组播组 G 的终端传输组播 IP 分组，将该组播 IP 分组作为分别以源终端 S 的 IP 地址和 RP 路由器的 IP 地址为源和目的地址的单播 IP 分组的净荷，并将该单播 IP 分组以单播传输方式传输给 RP 路由器，RP 路由器从该单播 IP 分组中分离出该组播 IP 分组，并将该组播 IP 分组通过共享组播树以组播传输方式传输给互连网络中属于组播组 G 的所有终端。如果源终端 S 需要向属于组播组 G 的终端传输大量组播 IP 分组，为了提高传输效率，可以构建（S，G）相关组播树，并直接通过（S，G）相关组播树完成源终端 S 向互连网络中属于组播组 G 的所有终端传输组播 IP 分组的过程。因此，PIM-SM 的工作过程如下：

- 向互连网络中的所有路由器公告 RP；
- 构建以 RP 为根，被所有活跃组播组共享的共享组播树；
- 如果需要，构建（S，G）相关组播树。

2. 消息类型和格式

1）消息类型

PIM-SM 的消息主要分为三类，一是用于将 RP 和组播组之间的关联信息扩散到互连网络中的所有路由器的消息，这一类消息主要有引导消息和 RP 通告消息，引导消息用于将 RP 和组播组之间的关联信息扩散到互连网络中的所有路由器，RP 通告消息用于在引导路由器中记录某个 RP 与一组组播组关联的信息。二是用于建立共享组播树和（S，G）相关组播树的消息，这一类消息主要有加入消息、剪枝消息等，加入消息用于在以 RP 或以源终端（或源网络）为根的组播树上增加某个分枝，剪枝消息在确定某个曾经活跃的组播组不再活跃后，剪去通往存在该曾经活跃的组播组的网络的分枝。三是在建立（S，G）相关的组播树前，用于实现源终端所在网络的指定路由器向 RP 传输组播 IP 分组的消息，这一类消息主要有注册消息和注册停止消息。注册消息用于实现源终端所在网络的指定路由器向 RP 传输组播 IP 分组的功能，注册停止消息用于 RP 要求某个指定路由器停止通过注册消息传输组播 IP 分组。

2）消息格式

引导消息由引导路由器（Bootstrap Router，BR）始发，以泛洪方式在互连网络中传播，到达互连网络中的所有路由器。如果互连网络中存在多个引导路由器，则通过竞争产生指定引导路由器，确定指定引导路由器的依据是优先级和引导路由器的 IP 地址。优先级高的引导路由器为指定引导路由器，如果存在多个具有相同最高优先级的引导路由器，其中 IP 地址最大的引导路由器为指定引导路由器。引导消息中给出互连网络中存在的所有 RP 及与每一个 RP 关联的组播组。PIM-SM 开始工作前，一是需要通过手工配置将若干个路由器作为引导路由器，并为这些路由器分配优先级；二是需要通过手工配置将若干个路由器作为 RP，并为每一个 RP 配置与其关联的一组组播组。

初始时，每一个引导路由器将自身作为指定引导路由器开始发送引导消息，在竞争出指定引导路由器后，由指定引导路由器周期性发送引导消息。当某个 RP 通过接收引导消息获得指定引导路由器地址后，通过 RP 通告消息将自己地址（RP 地址）及与其关联的一组组播组通告给指定引导路由器，指定引导路由器在以后发送的引导消息中将增加该 RP 及与

该 RP 关联的一组组播组。

允许互连网络中同时存在多棵以 RP 或源终端(或源网络)为根的组播树,每一个活跃组播组可以同时加入到这些以不同的 RP 或源终端(或源网络)为根的组播树。对于某个路由器,如果通往这些 RP 或源终端(或源网络)的传输路径有着相同的前一跳路由器,可以用一个加入消息完成向前一跳路由器发送有关增加分枝的信息。因此,加入消息中通过给出一组组播地址表明该分枝通往存在这一组活跃组播组的网络。加入消息中通过对应每一个组播组给出一组源终端地址(也可以是 RP 地址)表明该组播组可以同时加入以这些源终端或 RP 为根的组播树。

注册消息用于将组播 IP 分组封装成以源终端所在网络的指定路由器的 IP 地址为源 IP 地址,以 RP IP 地址为目的 IP 地址的单播 IP 分组,并通过互连网络实现该单播 IP 分组源终端所在网络的指定路由器至 RP 的传输过程。

大部分 PIM-SM 消息封装成 IP 分组后,源 IP 地址是发送该 PIM-SM 消息的接口的 IP 地址,目的 IP 地址是组播地址 224.0.0.13,表明接收端是该接口连接的网络中所有运行 PIM-SM 的路由器。PIM-SM 消息格式如图 7.22 所示。

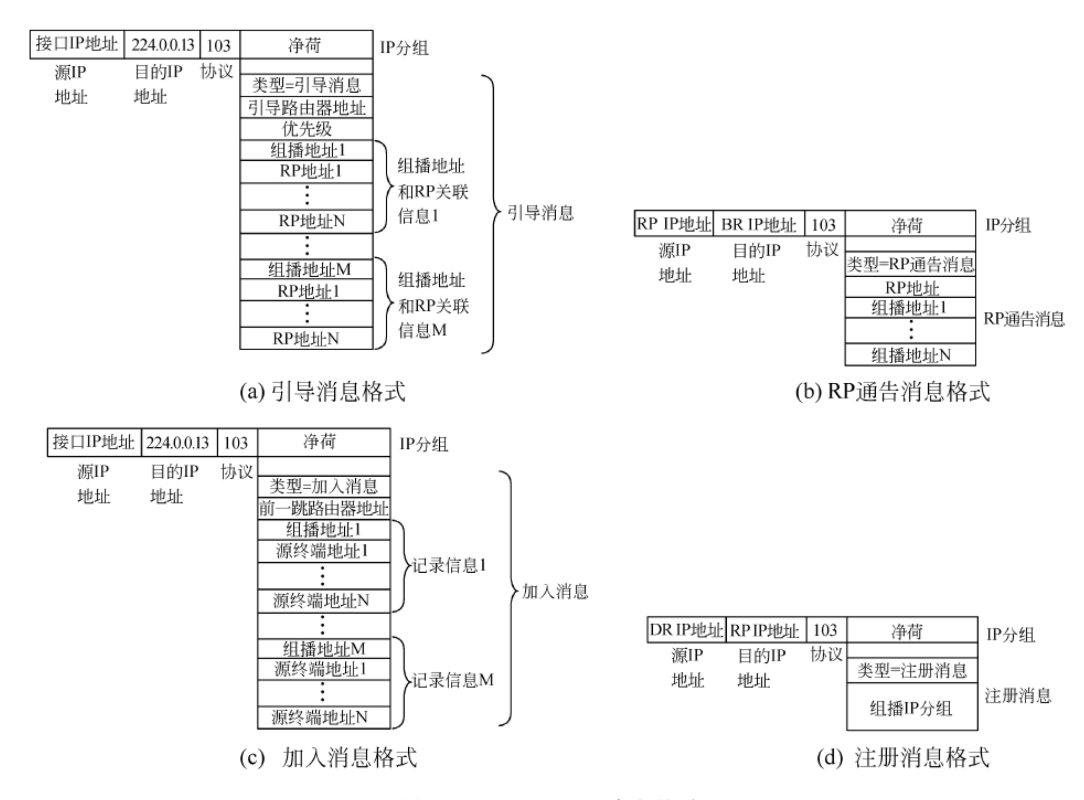


图 7.22 PIM-SM 消息格式

3. 通告 RP

引导路由器周期性地发送引导消息,初始引导消息中只包含引导路由器自身信息(IP 地址和优先级),引导路由器通过所有接口发送引导消息,引导消息的源 IP 地址是发送接口的 IP 地址,目的 IP 地址是组播地址 224.0.0.13。与 BR 相邻的所有路由器均接收到该引导消息,由于互连网络中所有其他路由器的单播路由表中已经建立用于指明通往 BR 的传

输路径的路由项,如表7.11所示的图7.23中除BR以外的所有路由器以BR为目的网络的单播路由项。每一个接收到引导消息的路由器首先用引导消息包含的BR地址检索单播路由表,找到匹配的单播路由项,如果该引导消息不是从该单播路由项的输出接口输入,表明该引导消息不是沿着BR至该路由器的最短路径到达,路由器将丢弃该引导消息。否则,存储该引导消息,并从除接收该引导消息以外的所有其他接口输出该引导消息。当然,从每一个接口输出的引导消息都以该接口的IP地址为源IP地址。当路由器R2通过接口3接收到引导消息,以BR地址为目的地址检索单播路由表,找到匹配的单播路由项＜BR,3,R1＞(表7.11中路由器R2以BR为目的网络的单播路由项),由于单播路由项中的输出接口是R2接收该引导消息的接口,路由器R2存储该引导消息,并通过接口1和接口4输出该引导消息。从路由器R2接口1输出的引导消息被路由器R4通过接口2接收,从路由器R2接口4输出的引导消息被路由器R5通过接口3接收,表7.11中路由器R4对应的以BR为目的网络的单播路由项,其输出接口是接口1,路由器R5对应的以BR为目的网络的单播路由项,其输出接口是接口2。由于路由器R4和路由器R5均发现接收该引导消息的接口不是该路由器以BR为目的网络的单播路由项的输出接口,因此路由器R4和路由器R5均丢弃该引导消息。图7.23给出到达所有路由器并被所有路由器存储的有效引导消息的泛洪过程。

表7.11 各个路由器以BR为目的网络的路由项

路由器名称	目的网络	输出接口	下一跳
R1	R3(BR)	3	直接
R2	R3(BR)	3	R1
R4(RP)	R3(BR)	1	直接
R5	R3(BR)	2	R4
R6	R3(BR)	4	直接
R7	R3(BR)	3	R6

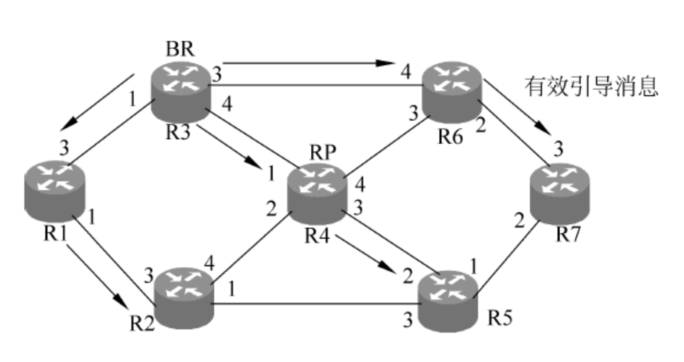


图7.23 引导消息泛洪过程

由于路由器R4已经被配置成RP,当路由器R4通过接收到的引导消息获知BR的地址,通过RP通告消息向BR通告自己的地址及关联的一组组播组。BR接收到RP发送的RP通告消息后,记录下RP地址及与该RP关联的一组组播组,并立即通过泛洪引导消息将这些信息通报给所有路由器。互连网络中的所有路由器记录下RP地址及与该RP关联的一组组播组。一旦直接连接的网络存在属于与该RP关联的某个组播组的终端,该路由

器将开始加入共享组播树的过程。

需要强调的是,某个组播组可能与若干个 RP 关联,必须保证所有路由器为特定组播组选择相同的 RP。

4. 构建共享组播树

当所有路由器接收到包含 RP 地址及与该 RP 关联的一组组播组的引导消息,且路由器 R2、路由器 R3、路由器 R6 和路由器 R7 均通过 IGMP 获知直接连接的网络中分别存在属于组播组 G1、组播组 G2、组播组 G1 和组播组 G2 的终端,这些路由器开始加入共享组播树的过程。开始加入共享组播树过程的路由器首先构建加入消息,加入消息中给出 RP 地址,并用 RP 地址检索单播路由表,找到匹配的单播路由项,将该单播路由项的下一跳路由器地址作为加入消息中的前一跳路由器地址,并通过该单播路由项的输出接口输出加入消息。对应表 7.12 所示的用于指明路由器 R7 通往 RP、源终端 S1 和源终端 S2 的传输路径的单播路由项,路由器 R7 构建的加入消息中给出 RP 地址——R4,前一跳路由器地址——R5 和组播地址——G2,并通过接口 2 输出该加入消息。当路由器 R5 通过接口 1 接收到该加入消息,在确定自己是加入消息中前一跳路由器地址指定的路由器后,在组播路由表中创建一项组播路由项,源终端(或源网络)字段值为 *,表明该组播路由项对应共享组播树,下游接口为接收该加入消息的接口——接口 1,然后,用 RP 地址检索表 7.13 所示的路由器 R5 单播路由表,找到匹配的单播路由项＜RP,2,直接＞,用单播路由项的输出接口——接口 2 作为该组播路由项的上游接口。路由器 R5 创建如表 7.14 所示的组播路由项。完成组播路由项创建后,路由器 R5 根据新的组播路由表构建加入消息,沿着通往 RP 的最短路径逐跳转发该加入消息。

表 7.12 路由器 R7 单播路由表

目的网络	输出接口	下一跳
RP	2	R5
S1	1	R6
S2	2	R5

表 7.13 路由器 R5 单播路由表

目的网络	输出接口	下一跳
RP	2	直接
S1	3	R2
S2	4	直接

表 7.14 路由器 R5 组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
*	G2	2	1

每一个路由器接收到请求加入(*,G2)的加入消息后,在组播路由表中检索与(*,G2)匹配的组播路由项,如果找到与(*,G2)匹配的组播路由项,将接收该加入消息的接口添加

到该组播路由项的下游接口列表中。如果没有找到与(*,G2)匹配的组播路由项,创建(*,G2)对应的组播路由项,将接收该加入消息的接口作为该组播路由项的下游接口,用RP地址检索单播路由表,找到匹配的单播路由项后,用该单播路由项的输出接口作为该组播路由项的上游接口。根据新的组播路由表创建加入消息,沿着通往RP的最短路径转发该加入消息。

当路由器R2、路由器R3、路由器R6和路由器R7始发的加入消息沿着通往RP的最短路径逐跳转发后到达RP,RP构建如表7.15所示的组播路由表,(*,G2)对应的组播路由项的下游接口列表包含接口1、接口2和接口3,(*,G1)对应的组播路由项的下游接口列表包含接口4。意味着在(*,G2)组播树中分别建立RP至路由器R2、路由器R3和路由器R7的分枝。在(*,G1)组播树中建立RP至路由器R6的分枝。

表7.15 RP组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
*	G2	—	1,2,3
*	G1	—	4

源终端S1发送的目的地址为组播地址G2的组播IP分组被源终端S1所在网络的指定路由器R1接收,由于指定路由器R1没有在组播路由表中找到与(S1,G2)匹配的组播路由项,由路由器R1将该组播IP分组封装成注册消息,并将注册消息封装成以路由器R1的IP地址为源IP地址,以RP路由器的IP地址为目的IP地址的单播IP分组,以单播传输方式将该单播IP分组传输给RP路由器。当RP路由器接收到注册消息,从中分离出源终端为S1、目的地址为组播地址G2的组播IP分组,开始该组播IP分组的组播过程。每一个路由器接收到该组播IP分组后,找出和(*,G2)匹配的组播路由项,确定该组播IP分组从该组播路由项的上游接口接收后,通过该组播路由项下游接口列表中给出的所有接口输出该组播IP分组,因此,该组播IP分组分别沿着RP路由器至路由器R2、路由器R3和路由器R7的分枝到达路由器R2、路由器R3和路由器R7。

对于图7.21所示的互连网络结构,为了维持(*,G2)和(*,G1)组播树,互连网络中(*,G2)和(*,G1)组播树经过的每一个路由器都需周期性地根据组播路由表生成并发送加入消息。

5. 构建源终端至RP路由器之间的分枝

源终端S1发送的目的IP地址为组播地址G2的组播IP分组需要封装成注册消息,并以单播传输方式完成注册消息源终端S1至RP路由器的传输过程,这样做会增加RP路由器的处理负担,降低链路带宽的效率,因此,当RP路由器接收到封装成注册消息的源终端S1发送的目的IP地址为组播地址G2的组播IP分组后,通过向源终端S1发送请求加入(S1,G2)组播树的加入消息,建立源终端S1至RP路由器的分枝。RP路由器始发的加入消息沿着RP路由器至源终端S1的最短路径逐跳转发,中间经过的路由器R2、路由器R1和RP路由器分别创建如表7.16、表7.17和7.18所示的(S1,G2)对应的组播路由项。这样,当源终端S1发送的目的IP地址为组播地址G2的组播IP分组到达路由器R1时,直接沿着路由器R1至RP路由器的分枝传输,中间经过路由器找出与(S1,G2)匹配的组播路由

项,确定该组播 IP 分组从该组播路由项的上游接口接收后,通过该组播路由项下游接口列表中给出的所有接口输出该组播 IP 分组。当该组播 IP 分组到达路由器 R2 时,路由器 R2 通过接口 2 和接口 4 输出该组播 IP 分组,使得该组播 IP 分组一方面沿着源终端至 RP 路由器分枝到达 RP 路由器,另一方面完成该组播 IP 分组源终端至其中一个目的终端的传输过程。注意,由于该组播 IP 分组直接通过源终端 S1 至路由器 R2 最短路径传输给路由器 R2,没有经过 RP 路由器,提高了链路效率。RP 路由器通过已经建立的共享组播树完成该组播 IP 分组 RP 路由器至其他目的终端的传输过程。

表 7.16 路由器 R1 组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
S1	G2	2	1

表 7.17 路由器 R2 组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
S1	G2	3	4,2
*	G2	4	2

表 7.18 RP 组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
*	G2	—	1,2,3
*	G1	—	4
S1	G2	2	1,3

值得强调的是,如果某个路由器的组播路由表中同时存在(*,G2)和(S1,G2)对应的组播路由项,源终端 S1 发送的目的 IP 地址为组播地址 G2 的组播 IP 分组同时匹配这两项组播路由项,但(S1,G2)对应的组播路由项优先于(*,G2)对应的组播路由项,路由器根据(S1,G2)对应的组播路由项转发该组播 IP 分组。

6. 构建(S1,G2)组播树分枝

当路由器 R3 和路由器 R7 接收到源终端 S1 发送的目的 IP 地址为组播地址 G2 的组播 IP 分组后,根据配置策略决定是否加入(S1,G2)组播树,在确定加入(S1,G2)组播树后,沿着路由器 R3 和路由器 R7 至源终端 S1 的最短路径逐跳转发请求加入(S1,G2)组播树的加入消息,加入消息中给出源终端 S1 地址和组播地址 G2。某个路由器接收到该加入消息后,在确定自己是加入消息中前一跳路由器地址指定的路由器的前提下,在组播路由表中检索与(S1,G2)匹配的组播路由项,如果找到与(S1,G2)匹配的组播路由项,将接收该加入消息的接口添加到该组播路由项的下游接口列表中。如果没有找到与(S1,G2)匹配的组播路由项,创建(S1,G2)对应的组播路由项,将接收该加入消息的接口作为该组播路由项的下游接口,用源终端 S1 地址检索单播路由表,找到匹配的单播路由项后,用该单播路由项的输出接口作为该组播路由项的上游接口。根据新的组播路由表创建加入消息,沿着通往源终端 S1 的最短路径转发该加入消息。

路由器 R3 始发的请求加入(S1,G2)组播树的加入消息到达路由器 R1。路由器 R7 始发的请求加入(S1,G2)组播树的加入消息到达路由器 R5,路由器 R5 将创建(S1,G2)对应的组播路由项,根据新的组播路由表创建加入消息,并向路由器 R2 发送该加入消息。由于路由器 R1 和路由器 R2 已经存在(S1,G2)对应的组播路由项,只是将接收该加入消息的接口添加到该组播路由项的下游接口列表中。完成(S1,G2)组播树构建后,路由器 R1、路由器 R2、路由器 R3、路由器 R5 和路由器 R7 的组播路由表分别如表 7.19～表 7.23 所示。

表 7.19 路由器 R1 组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
S1	G2	2	1,3

表 7.20 路由器 R2 组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
S1	G2	3	4,2,1
*	G2	4	2

表 7.21 路由器 R3 组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
S1	G2	1	2
*	G2	4	2

表 7.22 路由器 R5 组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
S1	G2	3	1
*	G2	2	1

表 7.23 路由器 R7 组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
S1	G2	2	1
*	G2	2	1

同样,为了维持(*,G2)和(S1,G2)组播树,(*,G2)和(S1,G2)组播树经过的所有路由器需要周期性地根据组播路由表生成并发送加入消息,如路由器 R7 生成的加入消息中,前一跳路由器地址为 R5,组播地址为 G2,该组播地址关联的源终端地址有两个,分别是 RP 地址和源终端 S1 地址,这是因为,路由器 R7 同时请求加入(*,G2)和(S1,G2)组播树,而且路由器 R7 通往 RP 路由器和源终端 S1 的最短路径有着相同的前一跳路由器——R5。

完成(S1,G2)组播树构建后,源终端 S1 发送的目的 IP 地址为组播地址 G2 的组播 IP 分组将分别沿着源终端 S1 至 RP 分枝、沿着源终端 S1 至路由器 R2、路由器 R3 和路由器 R7 分枝传输。沿着源终端 S1 至路由器 R2、路由器 R3 和路由器 R7 分枝传输的该组播 IP 分组分别通过接口 1 和接口 3 到达路由器 R3 和路由器 R5,由于这两个路由器的组播路由

表中存在与(S1,G2)匹配的组播路由项,且组播路由项的上游接口分别是这两个路由器接收该组播 IP 分组的接口,这两个路由器将通过组播路由项下游接口列表中给出的所有接口输出该组播 IP 分组。沿着源终端 S1 至 RP 分枝传输的该组播 IP 分组到达 RP 路由器后,沿着(*,G2)组播树传输,再次分别通过接口 4 和接口 2 到达路由器 R3 和路由器 R5,由于路由器 R3 和路由器 R5 组播路由表中与(S1,G2)匹配的组播路由项的上游接口分别是接口 1 和接口 3,路由器 R3 和路由器 R5 将丢弃该组播 IP 分组。显然,RP 路由器沿着(*,G2)组播树向路由器 R3 和路由器 R5 传输源终端 S1 发送的目的 IP 地址为组播地址 G2 的组播 IP 分组的过程是浪费的。

当路由器通过与(S1,G2)匹配的组播路由项的上游接口接收源终端 S1 发送的目的 IP 地址为组播地址 G2 的组播 IP 分组,且该路由器中与(*,G2)匹配的组播路由项的上游接口和与(S1,G2)匹配的组播路由项的上游接口不同时,该路由器将沿着该路由器至 RP 路由器的最短路径逐跳转发请求停止转发源终端 S1 发送的目的 IP 地址为组播地址 G2 的组播 IP 分组的剪枝消息。该路由器至 RP 路由器的最短路径经过的路由器将添加一项用于指明停止转发源终端 S1 发送目的 IP 地址为组播地址 G2 的组播 IP 分组的组播路由项。表 7.24 给出 RP 路由器接收到路由器 R3 和路由器 R5 始发的请求停止转发源终端 S1 发送目的 IP 地址为组播地址 G2 的组播 IP 分组的剪枝消息后的组播路由表内容。

表 7.24 RP 组播路由表

源终端(或源网络)	组播组	上游接口	下游接口列表
*	G2	—	1,2,3
*	G1	—	4
S1	G2	2	1p,3p

与(S1,G2)匹配的组播路由项的下游接口列表中后面紧跟字符 p 的接口是停止转发源终端 S1 发送的目的 IP 地址为组播地址 G2 的组播 IP 分组的接口。需要指出的是,为了维持上游路由器下游接口列表中每一个接口的状态,下游路由器必须周期性地发送加入消息或剪枝消息。

习题

7.1 什么是组播?组播的主要应用有哪些?

7.2 为什么需要构建广播树?用广播树传输组播 IP 分组和用泛洪方式传输组播 IP 分组有什么不同?以图 7.3 为例进行分析。

7.3 DVMRP 是何种类型的组播路由协议?它和 RIP 有哪些异同点?

7.4 根据图 7.24 所示互连网络结构和组播组分布,给出用 DVMRP 构建广播树的步骤和剪枝过程。

7.5 根据图 7.3 所示互连网络结构,给出用 DVMRP 构建路由器 R4 的组播路由表的过程。

7.6 IGMP 的作用是什么?

7.7 为什么需要嫁接过程?嫁接过程如何进行?

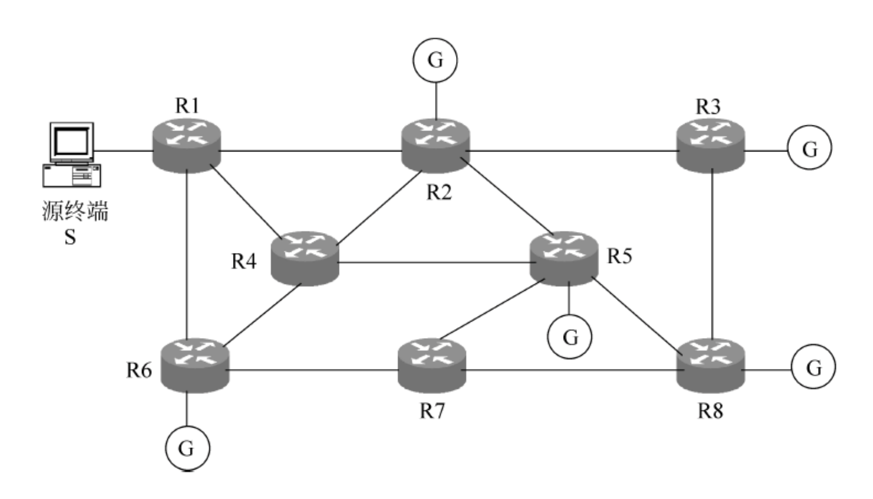


图 7.24 题 7.4 和题 7.9 图

7.8 下游路由器通过接收到嫁接应答消息确定上游路由器已经接收到嫁接消息，为什么不需要上游路由器发送剪枝应答消息？

7.9 根据图 7.24 所示互连网络结构和组播组分布，给出用 PIM-SM 构建（*，G）和（S，G）组播树的步骤。假定路由器 R4 为 RP 路由器。

7.10 DVMRP 和 PIM-SM 有什么本质不同？

7.11 针对图 7.25 所示的组播组分布，求出路由器 R5 对应源终端 192.1.2.1、组播组 G1、组播组 G2 和组播组 G3 的组播路由表。

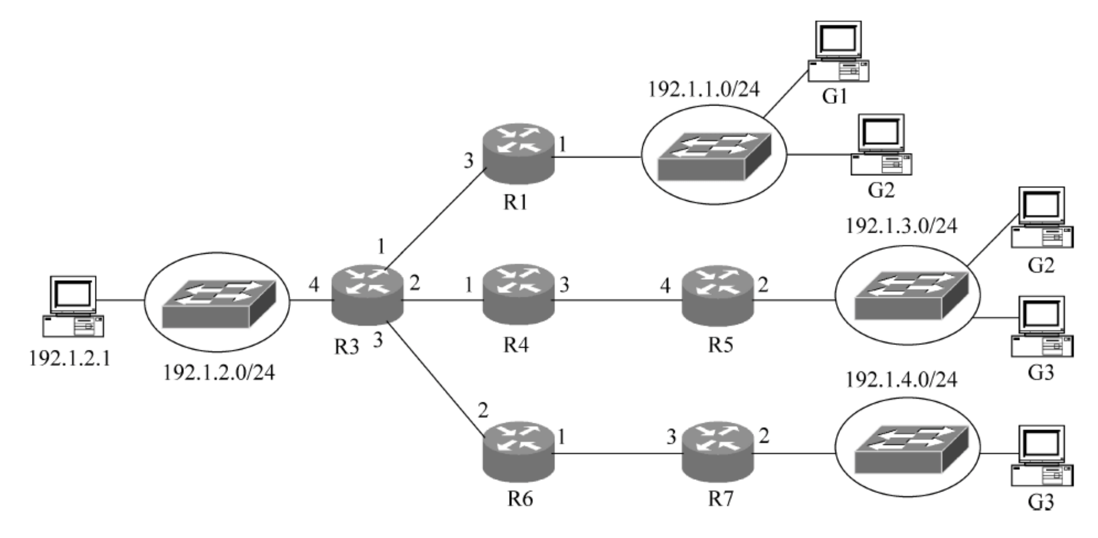


图 7.25 题 7.11 图

7.12 简述 PIM-SM 根据图 7.21 所示互连网络结构构建（S2，G2）组播树过程，并完善相关路由器的组播路由表。

7.13 简述单播路由表对 PIM-SM 完成组播路由表建立过程的重要性。

第8章 网络地址转换

出于安全和节省地址空间的需要，位于内部网络的终端(内部网络终端)只分配私有 IP 地址，但公共网络一般无法路由以私有 IP 地址为目的地址的 IP 分组，因此，分配私有 IP 地址的终端无法和位于公共网络中的终端(公共网络终端)通信。为了实现内部网络终端与公共网络终端之间通信，需要为某个有着与公共网络终端通信需求的内部网络终端分配一个公共网络能够识别的全球 IP 地址，且使得公共网络终端能够用该全球 IP 地址与该内部网络终端通信。这就要求内部网络终端在与公共网络终端的通信过程中，使用两个不同的 IP 地址，这两个不同 IP 地址分别是在内部网络使用的私有 IP 地址和在公共网络使用的全球 IP 地址，并由互连内部网络和公共网络的边界路由器实现这两个地址之间的转换，路由器实现这两个地址之间转换的技术称为网络地址转换技术。

8.1 NAT 基本概念

8.1.1 NAT 定义

如图 8.1 所示，由边界路由器 R 实现内部网络和外部网络的互连，但内部网络和外部网络本身可能是一个复杂的互连网络。由于受各种因素的限制，假定内部网络只能识别属于地址空间 192.168.3.0/24 和 172.16.3.0/24 的 IP 地址，外部网络只能识别属于地址空间 202.3.3.0/24 和 202.7.7.0/24 的 IP 地址。某个网络只能识别某个地址空间的含义是该网络中的路由器只能路由以属于该地址空间的 IP 地址为目的 IP 地址的 IP 分组。如果需要实现终端 A 与终端 B 之间通信，必须在内部网络为终端 B 分配一个属于地址空间 192.168.3.0/24 和 172.16.3.0/24 的 IP 地址，且内部网络能够将以该 IP 地址为目的 IP 地址的 IP 分组传输给边界路由器 R，边界路由器 R 能够将该 IP 分组转发给外部网络，并以终端 B 在外部网络中的地址作为该 IP 分组的目的 IP 地址。同样，必须在外部网络为终端 A 分配一个属于地址空间 202.3.3.0/24 和 202.7.7.0/24 的 IP 地址，且外部网络能够将以该 IP 地址为目的 IP 地址的 IP 分组传输给边界路由器 R，边界路由器 R 能够将该 IP 分组转发给内部网络，并以终端 A 在内部网络中的地址作为该 IP 分组的目的 IP 地址。这里假定为终端 B 在内部网络分配 IP 地址 172.16.3.7，为终端 A 在外部网络分配 IP 地址 202.7.7.3。这样终端 A 发送的、到达终端 B 的 IP 分组的源 IP 地址必须是外部网络分配给终端 A 的 IP 地址 202.7.7.3，终端 B 发送的、到达终端 A 的 IP 分组的源 IP 地址必须是

内部网络分配给终端 B 的 IP 地址 172.16.3.7。这就存在 4 个 IP 地址,终端 A 在内部网络使用的地址和终端 A 在外部网络使用的地址,终端 B 在内部网络使用的地址和终端 B 在外部网络使用的地址。通常将内部网络使用的地址称为本地地址(或私有地址),将外部网络使用的地址称为全球地址,因此,将位于内部网络的终端使用的本地地址称为内部本地地址,将位于内部网络的终端使用的全球地址称为内部全球地址,将位于外部网络的终端使用的本地地址称为外部本地地址,将位于外部网络的终端使用的全球地址称为外部全球地址。对于图 8.1 中的终端 A 和终端 B,这 4 个地址如表 8.1 所示。

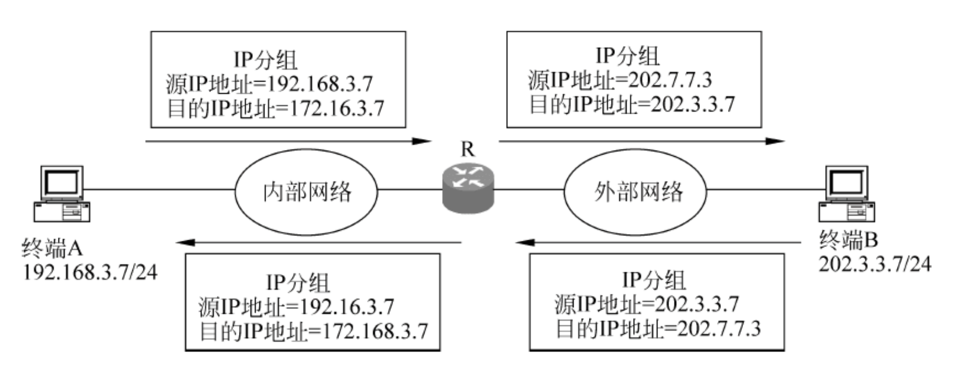


图 8.1 NAT 过程

表 8.1 终端 A 和终端 B 的本地和全球地址

内部本地地址（终端 A 内部网络地址）	内部全球地址（终端 A 外部网络地址）	外部本地地址（终端 B 内部网络地址）	外部全球地址（终端 B 外部网络地址）
192.168.3.7	202.7.7.3	172.16.3.7	202.3.3.7

边界路由器 R 的网络地址转换(Network address translation,NAT)技术就是一种对从内部网络转发到外部网络的 IP 分组实现源 IP 地址内部本地地址至内部全球地址的转换、目的 IP 地址外部本地地址至外部全球地址的转换,对从外部网络转发到内部网络的 IP 分组实现源 IP 地址外部全球地址至外部本地地址的转换、目的 IP 地址内部全球地址至内部本地地址的转换的技术。图 8.1 给出了终端 A 和终端 B 之间实现双向通信时,边界路由器 R 实现的地址转换过程。

8.1.2 私有地址空间

提出 NAT 的初衷是为了解决 IPv4 地址耗尽的问题,NAT 允许不同的内部网络分配相同的私有地址空间,且这些通过公共网络互连的、分配相同私有地址空间的内部网络之间可以实现相互通信。实现这一功能的前提是内部网络使用的私有地址空间和公共网络使用的全球地址空间之间不能重叠。为此,IETF 专门留出了三组 IP 地址作为内部网络使用的私有地址空间,公共网络使用的全球地址空间中不允许包含属于这三组 IP 地址的地址空间。这三组 IP 地址如下:

(1) 10.0.0.0/8。

(2) 172.16.0.0/12。

(3) 192.168.0.0/16。

多个内部网络允许使用相同的私有地址空间的原因是内部网络使用的私有地址空间对所有尝试与该内部网络通信的其他网络是不可见的,因此,两个使用相同私有地址空间的内部网络相互通信时,看到的都是对方经过转换后的全球 IP 地址。

8.1.3 NAT 应用

NAT 在互连网络设计中得到了广泛应用,以下是 NAT 常见的应用方式。

1. 局域网接入 Internet

目前家庭和小型企业接入 Internet 的过程如图 8.2 所示,接入控制设备对边界路由器进行身份鉴别,并对其分配全球 IP 地址。局域网内的终端分配私有 IP 地址,通过私有 IP 地址实现相互通信。如果局域网内的终端需要访问 Internet 中的资源,边界路由器需要完成终端私有 IP 地址与接入控制设备分配给它的全球 IP 地址之间的转换,当多个局域网内的终端同时访问 Internet 中的资源时,需要解决多个私有 IP 地址与单个全球 IP 地址之间的映射问题。

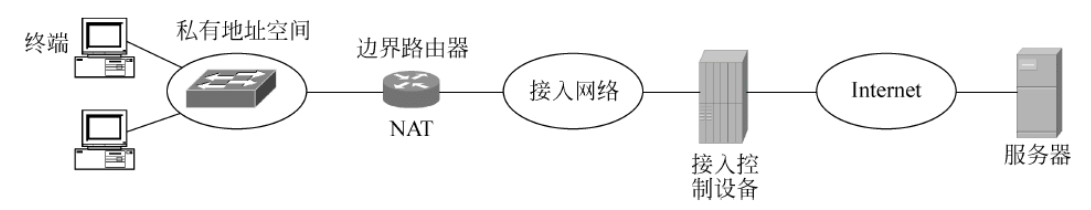


图 8.2 局域网接入 Internet 过程

2. 内部网络和外部网络互连

图 8.3 所示是实现企业网和 Internet 互连的互连网络结构,企业网使用私有 IP 地址空间,企业网内部终端通过私有 IP 地址实现相互通信。同时,企业网通过某个 Internet 服务提供者(Internet Service Provider,ISP)接入 Internet,ISP 分配给企业网一组全球 IP 地址,企业网内终端需要访问 Internet 中的资源时,必须使用 ISP 分配给企业网的一组全球 IP 地址,由路由器 R 完成企业网内终端私有 IP 地址与全球 IP 地址之间的转换。由于私有 IP 地址空间对 Internet 中的终端是透明的,因此,只能由企业网内部终端发起访问 Internet 中的资源的过程,Internet 中的终端不能主动发起访问企业网内部终端的过程。因此,图 8.3 所示的地址分配方式和互连网络结构使企业网具有一定的安全性。

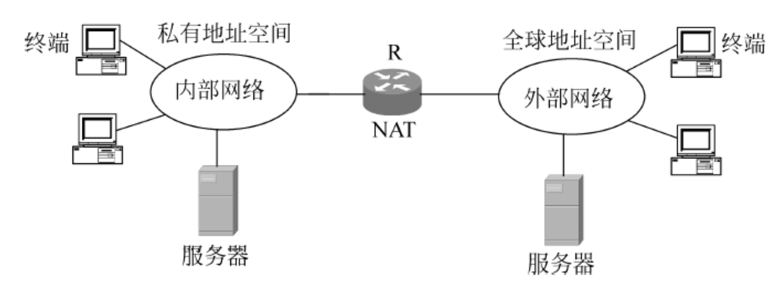


图 8.3 内部网络和外部网络互连

3. 内部网络之间相互通信

实现分别与外部网络互连的两个内部网络之间通信的机制有两种方式：一是通过虚拟专用网络（Virtual Private Network，VPN）技术；二是通过 NAT 技术。VPN 技术在图 8.4 所示的路由器 R1 和路由器 R2 之间建立隧道，对于被外部网络分隔的两个内部网络，隧道等同于点对点专用线路。这种情况下，外部网络对于这两个内部网络是透明的，整个互连网络完全等同于两个内部网络互连而成的大内部网络，两个内部网络必须分配不同的私有地址空间，连接在不同内部网络上的终端可以通过私有地址实现相互通信。NAT 技术允许两个内部网络分配相同的私有地址空间，但每一个内部网络使用的私有地址空间对另一个内部网络是透明的。因此，位于某个内部网络的终端，必须用全球 IP 地址与位于另一个内部网络中的终端通信。在实现终端 A 与终端 B 的通信过程中，终端 A 必须获知终端 B 对应的全球 IP 地址，构建以终端 A 私有 IP 地址为源 IP 地址、终端 B 全球 IP 地址为目的 IP 地址的 IP 分组。该 IP 分组经路由器 R1 转发后，源 IP 地址转换成终端 A 私有 IP 地址对应的全球 IP 地址，转换源 IP 地址后的 IP 分组经过外部网络到达路由器 R2，经路由器 R2 转发后，目的 IP 地址转换成终端 B 的私有 IP 地址。采用 NAT 技术的好处是不但能够实现连接在不同内部网络上的终端之间通信，还能够实现内部网络终端与外部网络终端之间的通信。

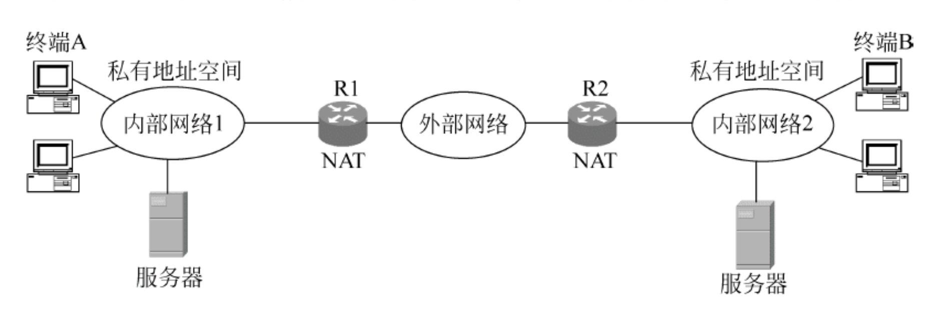


图 8.4　内部网络之间相互通信

4. 负载均衡

公共网站的访问量是很大的，单一服务器很难支撑如此大流量的访问，因此，公共网站常采用负载均衡技术。负载均衡技术用一组服务器来分担针对公共网站的访问流量，但公共网站提供给所有用户的是单个 IP 地址，因此，实现负载均衡的关键是将多个不同用户发送的、以公共网站 IP 地址为目的 IP 地址的 IP 分组均衡到多个不同的服务器上，同样，多个不同的服务器传输给不同用户的 IP 分组，到达用户时，统一以公共网站的 IP 地址为源 IP 地址。图 8.5 给出了通过 NAT 实现负载均衡的过程。路由器 R 能够将同一个全球 IP 地址轮流映射到多个不同的私有 IP 地址，使得以虚拟公共网站 IP 地址为目的 IP 地址的 IP 分组被传输给多个不同的服务器，反之，多个不同的服务器发送的 IP 分组，经路由器 R 转发后，有着相同的全球 IP 地址，即虚拟公共网站 IP 地址。

5. 多穴网络

有些企业网为了可靠性和传输效率，同时通过多个不同的 ISP 接入 Internet，通过对应的 ISP 访问 Internet 时，必须使用该 ISP 分配给企业网的全球 IP 地址，因此，同一个企业网

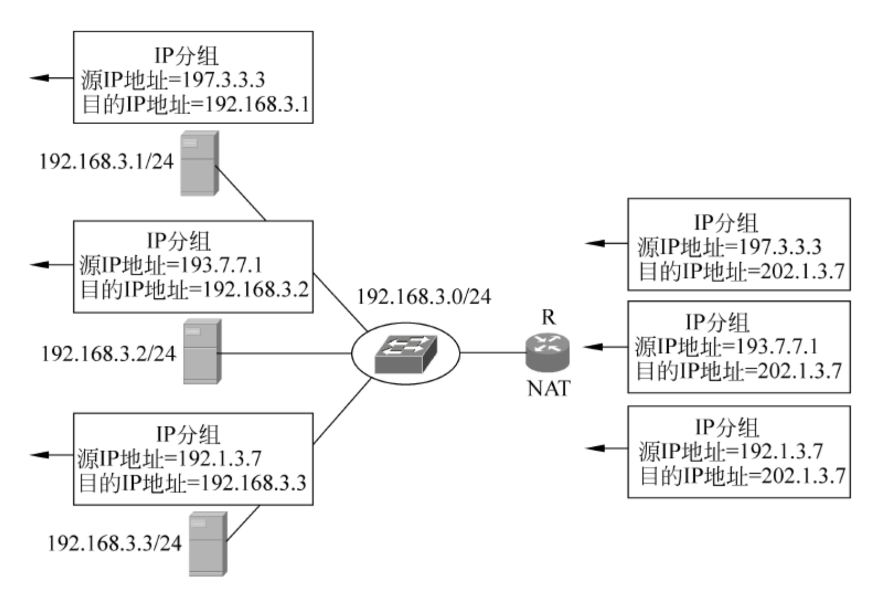


图 8.5 负载均衡实现过程

中的终端，传输给不同的 ISP 的 IP 分组的源 IP 地址是不同的。为了实现这一点，采用图 8.6 所示的多穴网络结构，企业网通过不同路由器连接不同的 ISP，企业网中的终端使用私有地址空间，但连接不同 ISP 的路由器将企业网中的终端发送的 IP 分组转发给对应的 ISP 时，将该 IP 分组的源 IP 地址转换成该 ISP 分配给企业网的全球 IP 地址。使得同一终端发送的 IP 分组，进入不同的 ISP 后，有着不同的源 IP 地址，且该 IP 地址就是该 ISP 分配给企业网的全球 IP 地址。

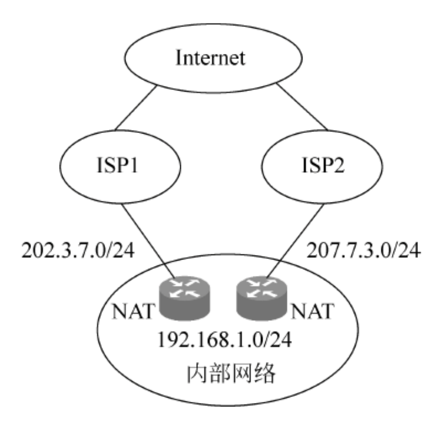


图 8.6 多穴网络结构

8.1.4 NAT 引发的问题

1. 重新计算检验和

每经过一跳路由器，TTL 字段值便会发生变化，因此 IP 分组首部中的首部检验和字段值需要重新计算。由于计算传输层首部中的检验和字段值时，IP 分组首部中的源和目的 IP 地址作为传输层的伪首部参与计算，在 IP 分组净荷是传输层报文的情况下，如果经过某个路由器时，IP 分组首部中的源和目的 IP 地址字段值发生变化，不仅需要重新计算 IP 分组首部中的首部检验和字段值，还需要重新计算传输层首部中的检验和字段值，这将大大增加该路由器转发 IP 分组时的计算负担，降低该路由器的 IP 分组转发速率。

2. 不便于分片

路由器完成网络地址转换时需要使用传输层首部中源和目的端口号字段值，如果某个封装了传输层报文的 IP 分组被分片，则只有第一片数据包含源和目的端口号字段值，如果包含第一片数据的 IP 分组首先到达需要进行网络地址转换的路由器，该路由器可以通过其

包含的源和目的端口号字段值完成网络地址转换，包含其他分片后数据的后续 IP 分组可以遵循包含第一片数据的 IP 分组完成网络地址转换时所建立的状态信息进行网络地址转换。由于 IP 分组不能保证按序传输，一旦出现包含其他分片后数据的 IP 分组先于包含第一片数据的 IP 分组到达需要进行网络地址转换的路由器的情况，由于该路由器无法对该 IP 分组进行网络地址转换，或者缓冲这样的 IP 分组，直到包含第一片数据的 IP 分组到达该路由器，但需要增加较大的缓冲器；或者丢弃该 IP 分组，导致所有包含分片后数据的 IP 分组都需重传。

3. 不利于数据加密

一旦 IP 分组净荷加密，路由器无法读到传输层首部中的源和目的端口号字段值，因而无法进行网络地址转换。多数情况下，NAT 与数据加密是对立的。

4. 需要增加应用层网关功能

有些应用层协议的协议数据单元(Protocol Data Unit，PDU)中包含源或目的终端的 IP 地址，如域名系统(Domain Name System，DNS)、文件传输协议(File Transfer Protocol，FTP)等，对于封装这种应用层协议的 PDU 的 IP 分组，不仅需要转换 IP 分组首部中的源和目的 IP 地址，还需转换应用层协议对应的 PDU 中包含的源或目的 IP 地址，由于不同应用层协议的 PDU 有着不同的格式和字段组成，因此，必须对应每一种应用层协议增加分析、处理对应的 PDU 的模块，这种用于分析、处理某个应用层协议对应的 PDU 的模块称为该应用层协议对应的应用层网关，路由器增加应用层网关，不仅增加成本，而且将大大增加路由器转发 IP 分组时的计算负担，降低路由器的 IP 分组转发速率。

5. 需要运行不同路由协议

内部网络的私有地址空间对外部网络是不可见的，因此，内部网络中的路由器不允许向外部网络中的路由器直接发送有关内部网络的路由信息，这就要求用相互独立的路由协议分别产生用于指明通往内部网络中各个子网的传输路径的路由项与用于指明通往外部网络中各个子网的传输路径的路由项。

8.2 NAT 工作过程

8.2.1 NAT 分类

根据完成 NAT 需要涉及的协议层，可以分为涉及网络层 NAT、涉及传输层 NAT 和涉及应用层 NAT，涉及网络层 NAT 实现私有 IP 地址空间与全球 IP 地址空间之间的双向映射，涉及传输层 NAT 完成私有 IP 地址空间与传输层端口号之间的双向映射，涉及应用层 NAT 需要同步修改应用层 PDU 中的源或目的 IP 地址。涉及网络层的 NAT 目前有基本 NAT，简称 NAT；涉及传输层的 NAT，目前有端口地址转换(Port Address Translation，PAT)；涉及应用层的 NAT，目前有应用层网关(Application Level Gateway，ALG)。NAT 和 PAT 可以动态或是手工配置建立私有 IP 地址与全球 IP 地址之间映射、私有 IP 地址空

间与传输层端口号之间映射，因此NAT和PAT分为动态NAT、动态PAT和静态NAT、静态PAT。

8.2.2 PAT

1. 动态PAT

当图8.7中分配了本地IP地址的终端想访问Internet中的服务器(192.1.2.5)时，就构建一个以本地IP地址(192.168.1.1)为源IP地址，服务器IP地址(192.1.2.5)为目的IP地址的IP分组。由于配置终端时，默认网关地址为192.168.1.254，终端将这样的IP分组发送给边界路由器。分配给终端的本地IP地址只在内部网络内有效，Internet并不认可这种地址分配，如果服务器以此地址作为目的IP地址向内部网络内终端发送IP分组的话，Internet是无法正确地将该IP分组转发给内部网络内终端的，因此，须用ISP分配给边界路由器的全球IP地址作为IP分组的源IP地址。但由于ISP分配给边界路由器的全球IP地址只有一个，如果同时有多个内部网络内的终端访问Internet的话，这些内部网络内的终端用于访问Internet的IP分组经过边界路由器转发后，就有了相同的源IP地址(192.1.1.1)，而服务器回复给这些内部网络内的终端的IP分组的目的IP地址都是相同的，边界路由器如何能够从这些目的IP地址都相同的IP分组中鉴别出属于不同内部网络内终端的IP分组呢?

IP地址是网络层地址，只能唯一标识网络终端，而通信是进程间的事情，对于多任务系统，终端上可能同时运行多个进程，因此，必须在传输层报文首部提供用于唯一标识进程的端口号。这样，标识IP分组发送实体的信息由两部分组成：源IP地址和源端口号，在无法用源IP地址唯一标识源终端的情况下，可用源端口号来唯一标识源终端。但源终端传输层进程构建传输层报文时，只是用源端口号唯一标识终端内的发送进程，源端口号具有本地意义，即不同的终端可能用相同的源端口号标识终端内的进程。因此，边界路由器必须用内部网络内唯一的源端口号取代IP分组中原来的源端口号，以此实现用源端口号唯一标识内部网络内终端的目的。这种通过将内部网络内不同终端映射到不同源端口号的方法就是端口地址转换(Port Address Translation，PAT)。边界路由器在用ISP分配给它的全球IP地址取代IP分组中的源IP地址时，必须用内部网络内唯一的源端口号取代IP分组中原来的源端口号，然后在地址转换表中记录一项，把IP分组原来的源端口号、源IP地址和边界路由器取代的唯一的源端口号和全球IP地址绑定在一起。当服务器回送的IP分组到达边界路由器时，用该IP分组的目的端口号去检索地址转换表，找到对应项，用对应项中的源IP地址、原来的源端口号取代该IP分组的目的IP地址、目的端口号，然后将取代后的IP分组转发给局域网，如图8.7所示。

两个进程间的通信过程称为会话，在会话期间，必须采用相同的地址转换过程，即属于同一会话的IP分组，转换后的源IP地址和源端口号必须相同，因此，必须将图8.7所示的地址转换表中的每一项和某个会话绑定在一起，在该会话开始时创建该转换项，在会话结束时删除该转换项。每一个会话用源和目的IP地址、源和目的端口号唯一标识。

2. 静态PAT

图8.7所示的地址转换表在边界路由器接收到局域网中终端发送的属于某个特定会话

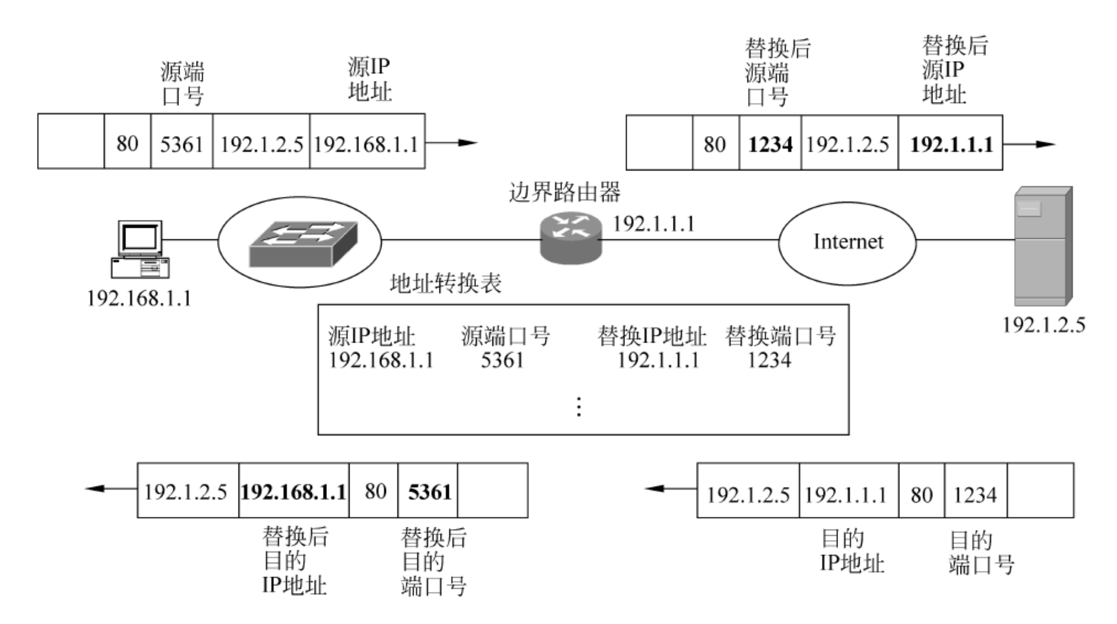


图 8.7　端口地址转换方法实现地址转换的过程

的第一个 IP 分组时创建，如局域网内终端发送的请求与 Internet 中某个服务器建立 TCP 连接的 TCP 连接请求报文。只有在边界路由器建立了与某个会话绑定的内部网络的本地地址与局域网内唯一的端口号之间映射，Internet 中的服务器才能与内部网络中分配了该本地地址的终端通信。如果局域网中的服务器向 Internet 中的终端开放，即允许 Internet 中的终端发起访问内部网络中的服务器的过程，需要静态配置服务器本地地址与局域网内唯一端口号之间的映射，这种通过手工配置建立某个本地地址与局域网内唯一端口号之间映射的机制称为静态 PAT。

静态 PAT 通过手工配置建立如图 8.8 所示的地址转换表，边界路由器如果从连接外部网络(Internet)的接口接收到 IP 分组，在地址转换表中检索全球地址和全球端口号与 IP 分组的目的 IP 地址和目的端口号匹配的地址转换项，用该地址转换项中的本地地址和本地端口号取代 IP 分组中的目的 IP 地址和目的端口号。边界路由器如果从连接内部网络的接口

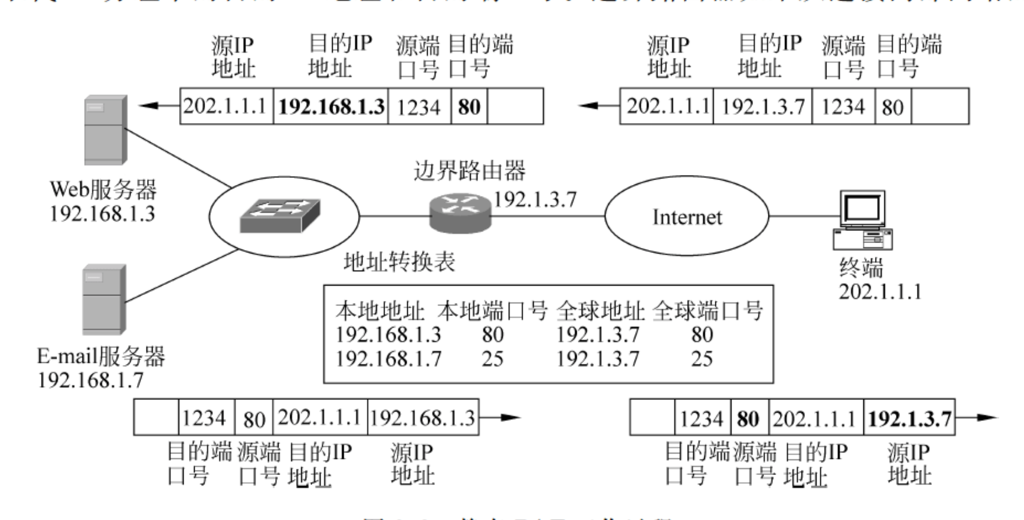


图 8.8　静态 PAT 工作过程

接收到 IP 分组，在地址转换表中检索本地地址和本地端口号与 IP 分组的源 IP 地址和源端口号匹配的地址转换项，用该地址转换项中的全球地址和全球端口号取代 IP 分组中的源 IP 地址和源端口号。通过静态 PAT，图 8.8 中连接在 Internet 中的终端可以发起访问内部网络中 Web 服务器的过程。IP 分组端口地址转换过程如图 8.8 所示。

8.2.3 NAT

1. 动态 NAT

动态 NAT 用于动态建立内部网络本地地址与全球地址之间的映射，和端口地址转换不同，动态 NAT 需要分配给内部网络一组全球 IP 地址，而不是一个全球 IP 地址，所有需要访问 Internet 的终端必须先建立该终端本地地址与某个全球地址之间的映射。

实现动态 NAT，首先需要定义全球 IP 地址池，如图 8.9 中定义的全球 IP 地址池：192.1.1.2～192.1.1.5，然后，需要定义允许和全球 IP 地址池中全球 IP 地址建立映射的本地地址范围。完成这些定义后，当某个分配了本地地址的终端发起访问 Internet 过程时，该终端发送以分配给该终端的本地地址为源 IP 地址的 IP 分组，路由器通过连接内部网络的接口接收到该 IP 分组后，如果在地址转换表中检索不到本地地址与该 IP 分组的源 IP 地址相同的地址转换项，路由器在全球 IP 地址池中选择一个未分配的全球 IP 地址，在地址转换表中创建本地内部地址为该 IP 分组的源 IP 地址、内部全球地址为全球 IP 地址池中选择的全球 IP 地址的地址转换项，并用内部全球 IP 地址取代该 IP 分组的内部本地地址。如果全球 IP 地址池中的全球 IP 地址已经分配完毕，路由器将丢弃该 IP 分组。如果路由器通过连接外部网络的接口接收到 IP 分组，在地址转换表中检索内部全球地址与该 IP 分组的目的 IP 地址相同的地址转换项，并用该地址转换项给出的内部本地地址取代该 IP 分组的目的地址。如果地址转换表中检索不到内部全球地址与该 IP 分组的目的 IP 地址相同的地址转换项，路由器丢弃该 IP 分组。

如图 8.9 所示，当本地地址为 192.168.1.1 的内部网络终端发送用于访问 Internet 中资源的第一个 IP 分组时，路由器从还没有分配的全球 IP 地址中选择一个全球 IP 地址(192.1.1.2)分配给该终端，并创建内部本地地址为 192.168.1.1、内部全球地址为 192.1.1.2 的地址转换项。以后，所有通过路由器连接内部网络接口接收到的源 IP 地址为内部本地地址 192.168.1.1的 IP 分组，源 IP 地址一律用内部全球地址 192.1.1.2 替代。同样，路由器一旦通过连接 Internet 的接口接收到目的 IP 地址为 192.1.1.2 的 IP 分组，用内部本地地址 192.168.1.1 取代该 IP 分组的目的 IP 地址。

地址转换表中的每一项地址转换项都关联一个定时器，每当通过路由器连接内部网络的接口接收到源 IP 地址为该地址转换项中内部本地地址的 IP 分组，刷新与该地址转换项关联的定时器，一旦关联的定时器溢出，将删除该地址转换项，路由器可以重新分配该地址转换项中的内部全球 IP 地址。

2. 静态 NAT

动态 NAT 都只能实现单向会话，即会话发起者必须是内部网络中的终端，由内部网络终端发送用于访问 Internet 中资源的第一个 IP 分组，并由该 IP 分组在内部网络和外部网

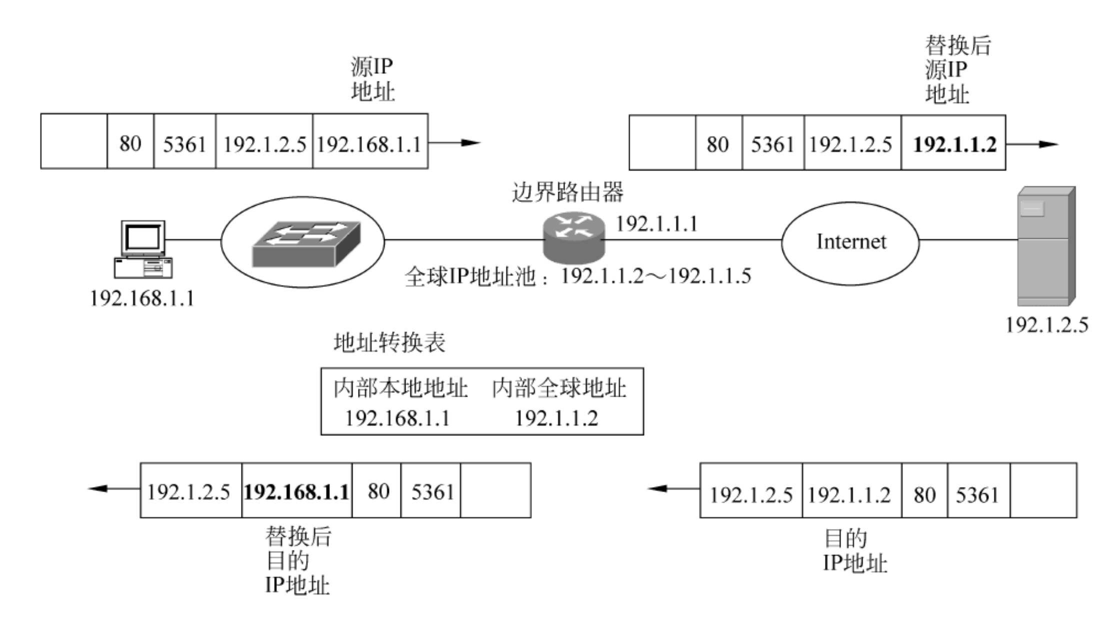


图 8.9　动态 NAT 方法实现地址转换的过程

络之间的边界路由器建立内部本地地址与内部全球地址之间的映射。如果需要由 Internet 中的终端发起访问内部网络中资源的过程，由于在边界路由器建立内部本地地址与内部全球地址之间的映射前，Internet 中的终端无法通过全球地址来唯一标识某个内部网络终端，因而无法向内部网络终端发送 IP 分组。因此，如果想要实现双向会话，需要手工建立某个本地地址与某个全球地址之间的映射，这样，Internet 中的终端可以用该全球地址访问内部网络中分配了该本地地址的终端。这种通过手工配置建立某个本地地址与某个全球地址之间的映射机制称为静态 NAT。

如图 8.10 所示，通过手工配置建立内部本地地址 192.168.1.1 与内部全球地址192.1.1.2 之间的映射，地址转换表中长期存在用于表明该映射的地址转换项。当路由器通过连接 Internet 的接口接收到以全球地址 192.1.1.2 为目的 IP 地址的 IP 分组，在地址转换表中检索到内部全球地址为 192.1.1.2 的地址转换项，用该地址转换项中的内部本地地址 192.168.1.1取代该 IP 分组的目的 IP 地址。

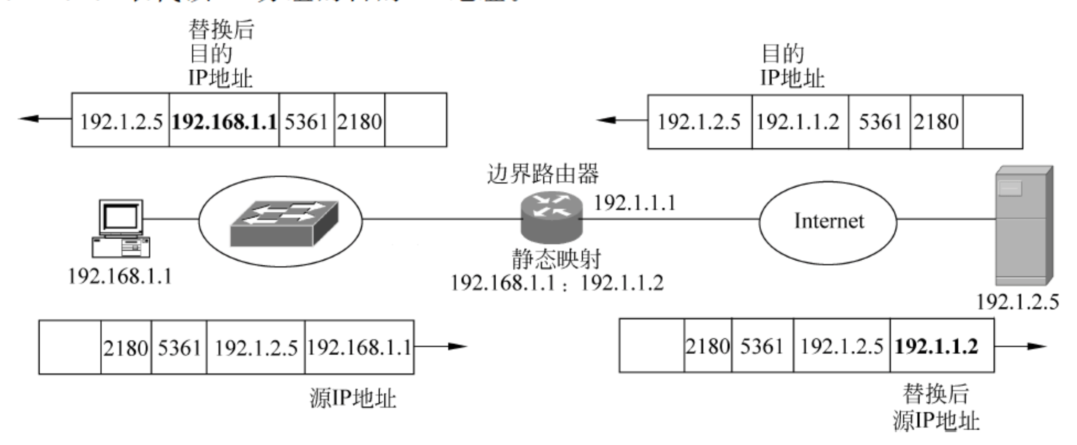


图 8.10　静态 NAT 方法实现地址转换的过程

当路由器通过连接内部网络的接口接收到以本地地址 192.168.1.1 为源 IP 地址的 IP 分组，在地址转换表中检索到内部本地地址为 192.168.1.1 的地址转换项，用该地址转换项中的内部全球地址 192.1.1.2 取代该 IP 分组的源 IP 地址。

8.2.4 应用层网关

1. 功能说明

NAT 根据已经建立的内部本地地址与内部全球地址之间的映射，完成 IP 分组源 IP 地址本地地址至全球地址转换，或者目的 IP 地址全球地址至本地地址转换，地址转换发生网络层，只需修改 IP 分组首部。PAT 根据建立的本地地址、本地端口号与全球地址、全球端口号之间的映射，完成 IP 分组及作为 IP 分组净荷的传输层报文源 IP 地址本地地址至全球地址、源端口号本地端口号至全球端口号的转换，或者目的 IP 地址全球地址至本地地址、目的端口全球端口号至本地端口号转换，地址转换发生在网络层和传输层，需要同步修改 IP 分组首部和传输层报文首部。

某些应用层协议，如 FTP、DNS 等，PDU 中包含源或目的终端的 IP 地址、源或目的进程对应的端口号，对于以这种类型应用层 PDU 为净荷的传输层报文和 IP 分组，仅仅完成 IP 分组首部和传输层报文首部同步修改是不够的，还需同步修改对应的应用层 PDU，这种通过分析某个应用层协议对应的 PDU，实现对该 PDU 与 IP 分组首部、传输层首部同步修改的地址转换技术称为应用层网关。其实，应用层网关是一种广义定义，所有用于分析、处理某个应用层协议对应的 PDU 的模块称为该应用层协议对应的应用层网关。因此，应用层网关是应层协议相关的，同样，作为一种 NAT 类型的应用层网关也是应用层协议相关的，这里，主要讨论 FTP 相关的应用层网关的工作过程。

2. FTP 应用层网关工作过程

下面以图 8.11 所示的终端 A 发起访问 FTP 服务器过程为例，讨论 FTP 应用层网关的工作过程。

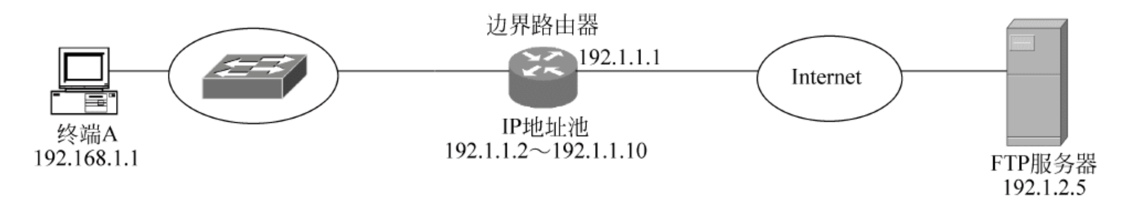


图 8.11 实现 ALG 网络结构

终端 A 访问 FTP 服务器涉及的信息交换过程如图 8.12 所示，首先由终端 A 发起建立与 FTP 服务器之间的控制 TCP 连接，边界路由器接收到终端 A 发送的请求建立控制 TCP 连接的请求报文后，建立终端 A 本地地址 192.168.1.1 与全球地址 192.1.1.2 之间的映射，FTP 服务器通过全球地址 192.1.1.2 实现与终端 A 通信。

当终端 A 请求从 FTP 服务器下载文件时，需要建立终端 A 与 FTP 服务器之间的数据 TCP 连接。FTP 服务器进程启动后，FTP 服务器进程被动侦听 TCP 端口号 21，等待 FTP

客户端向其发送请求建立控制 TCP 连接的请求报文，因此，FTP 控制 TCP 连接是由 FTP 客户端发起建立的，但 FTP 数据 TCP 连接是由 FTP 服务器发起建立的，因此，建立终端 A 与 FTP 服务器之间的数据 TCP 连接前，终端 A 必须被动侦听某个 TCP 端口号，并把终端 A 的 IP 地址与侦听的 TCP 端口号通过命令发送给 FTP 服务器，但终端 A 通过 FTP 命令给出的是终端 A 的本地地址 192.168.1.1，如果 FTP 服务器发起建立以终端 A 通过命令给出的 IP 地址和端口号为目的 IP 地址和目的端口号的 TCP 连接，该次 TCP 连接建立过程注定是要失败的。为了使由 FTP 服务器发起的 FTP 服务器与终端 A 之间的数据 TCP 连接建立过程能够成功进行，边界路由器必须根据地址转换表中已经建立的终端 A 本地地址 192.168.1.1 与全球地址 192.1.1.2 之间的映射，同步修改终端 A 通过命令给出的终端 A 本地地址 192.168.1.1。这就要求边界路由器能够分析 FTP PDU，识别不同的 FTP 命令及这些命令携带的参数，检测出用于给出数据 TCP 连接一端插口地址（IP 地址和 TCP 端口号）的命令，并对作为命令参数的 IP 地址做出同步修改。

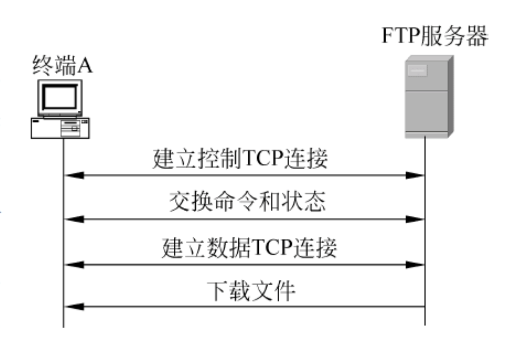


图 8.12 FTP 下载文件过程

不同应用层协议的信息交换过程是不同的，PDU 格式和内容也是不同的，因此，必须针对每一种应用层协议配置对应的应用层网关。应用层网关根据已经建立的本地地址与全球地址之间的映射同步修改应用层 PDU 的过程是非常复杂的，因此，会严重影响路由器转发 IP 分组的速率。

8.3 NAT 应用方式

8.3.1 双穴网络结构

1. 基本情况

双穴网络结构如图 8.13 所示，内部网络 192.168.1.0/24 通过两个 ISP 接入 Internet，

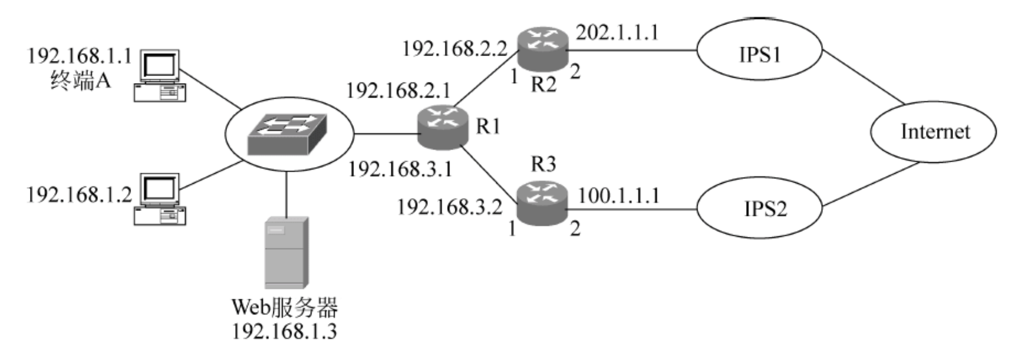


图 8.13 双穴网络结构

两个 ISP 分配给内部网络的全球地址分别是 202.1.1.1 和 100.1.1.1，内部网络通过动态 PAT 实现内部网络终端访问 Internet 服务器的过程，通过静态 PAT 实现 Internet 中的终端访问内部网络 Web 服务器的过程。为了均衡负载，同时也为了提高传输效率，目的地址属于 ISP1 的 IP 分组转发给 ISP1，目的地址属于 ISP2 的 IP 分组转发给 ISP2。ISP1 和 ISP2 的全球地址块分别是 202.1.0.0/16 和 100.1.0.0/16。其他 IP 分组转发给 ISP1。

2. 基本配置

1）路由器路由表（表 8.2～表 8.4）

路由器 R1 路由表保证将目的地址属于 ISP1 的 IP 分组转发给 ISP1、将目的地址属于 ISP2 的 IP 分组转发给 ISP2，将其他目的地址的 IP 分组转发给 ISP1。路由器 R2 和路由器 R3 路由表将目的地址为外部网络地址的 IP 分组转发给 ISP1 或 ISP2，下一跳地址应该分别是 ISP1 或 ISP2 中与路由器 R2 或路由器 R3 直接连接的路由器的地址。将目的地址为本地地址的 IP 分组转发给路由器 R1。

表 8.2　路由器 R1 路由表

目的网络	下一跳
192.168.1.0/24	直接
202.1.0.0/16	192.168.2.2
100.1.0.0/16	192.168.3.2
0.0.0.0/0	192.168.2.2

表 8.3　路由器 R2 路由表

目的网络	下一跳
192.168.1.0/24	192.168.2.1
0.0.0.0/0	ISP1 中与 R2 直接相连的路由器

表 8.4　路由器 R3 路由表

目的网络	下一跳
192.168.1.0/24	192.168.3.1
0.0.0.0/0	ISP2 中与 R3 直接相连的路由器

2）边界路由器 R2 和路由器 R3 的 PAT 配置

①指定边界路由器连接内部网络的接口和连接外部网络的接口。②指定使用动态 PAT，并指定内部网络允许访问外部网络中资源的本地 IP 地址范围 192.168.1.0/24。③在路由器 R2 中配置静态 PAT 映射 192.168.1.3:80 与 202.1.1.1:80，在路由器 R3 中配置静态 PAT 映射 192.168.1.3:80 与 100.1.1.1:80。这样，Internet 中的终端可以分别通过插口地址 202.1.1.1:80 和 100.1.1.1:80 访问内部网络中的 Web 服务器。

3. IP 分组传输过程

以内部网络中的终端 A 访问外部网络中某个 IP 地址为 202.1.7.7 的服务器为例，讨论 IP 分组内部网络与外部网络之间的双向传输过程。内部网络中的终端 A 向 ISP1 中 IP

地址为 202.1.7.7 的服务器传输 IP 分组的过程如下。

- 终端 A 构建以 192.168.1.1 为源 IP 地址、202.1.7.7 为目的 IP 地址的 IP 分组，根据配置的默认网关地址将该 IP 分组传输给路由器 R1，路由器 R1 用该 IP 分组的目的 IP 地址检索路由表，找到匹配的路由项，并将该 IP 分组传输给该路由项指定的下一跳路由器——路由器 R2；
- 路由器 R2 确定通过连接内部网络的接口接收到该 IP 分组，用该 IP 分组的目的 IP 地址检索路由表，找到匹配的路由项，发现该路由项中的输出接口是连接外部网络的接口，并且该 IP 分组的源 IP 地址属于允许进行 PAT 操作的本地地址范围(192.168.1.1∈192.168.1.0/24)，对该 IP 分组实施 PAT 操作，用全球地址 202.1.1.1作为该 IP 分组的源 IP 地址，产生一个内部网络唯一的端口号，用该端口号作为传输层报文的源端口号，在地址转换表中创建一项用于建立本地地址、原来的源端口号(本地端口号)与全球地址、内部网络唯一的端口号(全球端口号)之间映射的地址转换项；
- 路由器 R2 将完成 PAT 操作后的 IP 分组转发给 ISP1，经过 ISP1 中路由器的逐跳转发，该 IP 分组到达 IP 地址为 202.1.7.7 的服务器。

ISP1 中 IP 地址为 202.1.7.7 的服务器向内部网络中的终端 A 传输 IP 分组的过程如下。

- 服务器构建以 IP 地址 202.1.7.7 为源 IP 地址、以全球地址 202.1.1.1 为目的 IP 地址的 IP 分组，以全球端口号作为 IP 分组封装的传输层报文的目的端口号，该 IP 分组经过 ISP1 中路由器的逐跳转发，到达路由器 R2；
- 路由器 R2 确定通过连接外部网络的接口接收到该 IP 分组，在地址转换表中检索全球地址等于该 IP 分组的目的 IP 地址、全球端口号等于该 IP 分组封装的传输层报文的目的端口号的地址转换项，找到匹配的地址转换项后，用该地址转换项中的本地地址作为该 IP 分组的目的 IP 地址，本地端口号作为该 IP 分组封装的传输层报文的目的端口号；
- 用完成 PAT 操作后的 IP 分组的目的 IP 地址 192.168.1.1 检索路由器 R2 路由表，确定下一跳路由器 R1，将该 IP 分组传输给路由器 R1。该 IP 分组经路由器 R1 转发后，到达终端 A。

需要指出的是，如果路由器从连接内部网络的接口接收到 IP 分组，首先通过检索路由表确定该 IP 分组的输出接口，在确定输出接口是连接外部网络的接口，且 IP 分组的源 IP 地址属于允许进行地址转换的内部网络地址范围时，才开始进行地址转换过程。如果路由器从连接外部网络的接口接收到 IP 分组，首先检索地址转换表，如果在地址转换表中找到全球地址与该 IP 分组的目的 IP 地址相同、全球端口号与该 IP 分组封装的传输层报文的目的端口号相同的地址转换项，先进行地址转换过程，然后检索路由表。如果在地址转换表中找不到与该 IP 分组匹配的地址转换项，则直接检索路由表。

8.3.2 实现内部网络和外部网络通信

1. 基本情况

网络结构如图 8.14 所示，内部网络分配私有地址块 192.168.1.0/24，路由器 R2 只能

路由以全球 IP 地址为源和目的 IP 地址的 IP 分组，因此，需要为内部网络分配全球 IP 地址块 193.1.1.16 /28，路由器 R2 中必须建立用于指明通往目的网络 193.1.1.16/28 的传输路径的路由项。当内部网络中的终端访问外部网络中的终端或服务器时，需由路由器 R1 完成私有地址与全球 IP 地址之间的转换，并建立地址转换表。

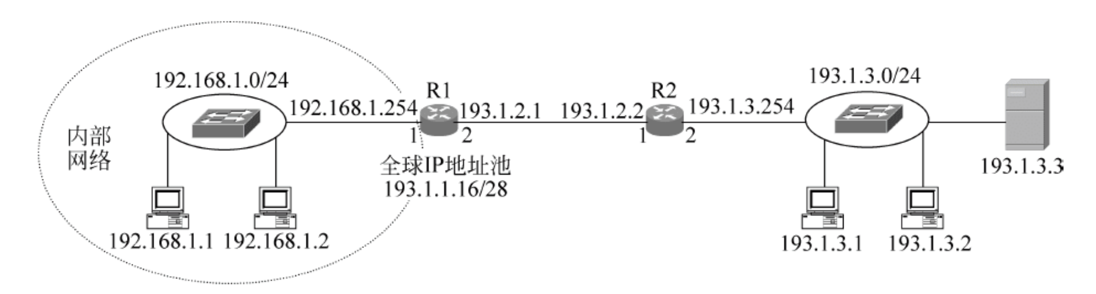


图 8.14 实现动态 NAT 的网络结构

2. 基本配置

1）路由器路由表（表 8.5、表 8.6）

这里只给出用于指明通往末端网络的传输路径的路由项，路由器 R1 分别给出用于指明通往末端网络 192.168.1.0/24 和 193.1.3.0/24 的传输路径的路由项。内部网络 192.168.1.0/24 对路由器 R2 是透明的，但路由器 R2 需要把目的 IP 地址属于网络地址 193.1.1.16/28 的 IP 分组转发给路由器 R1。

表 8.5 路由器 R1 路由表

目的网络	下一跳
192.168.1.0/24	直接
193.1.3.0/24	193.1.2.2

表 8.6 路由器 R2 路由表

目的网络	下一跳
193.1.3.0/24	直接
193.1.1.16/28	193.1.2.1

2）路由器 R1 动态 NAT 配置

①配置全球 IP 地址池 193.1.1.16/28。②配置路由器连接内部网络和外部网络接口。③指定使用动态 NAT，并配置允许进行地址转换的内部网络本地地址（私有地址）范围。

图 8.14 所示的内部网络与外部网络之间的 IP 分组传输过程和图 8.13 所示的内部网络与外部网络之间的 IP 分组传输过程基本相同，不同的是路由器 R1 创建的地址转换项只需建立内部网络本地地址与全球 IP 地址池中某个未分配的全球 IP 地址之间的映射，路由器 R1 对于通过连接内部网络的接口接收到的 IP 分组，完成该 IP 分组源 IP 地址本地地址至全球地址的转换，对于通过连接外部网络的接口接收到的 IP 分组，完成该 IP 分组目的 IP 地址全球地址至本地地址的转换。需要指出的是，进行 PAT 操作的 IP 分组的净荷必须是传输层报文，但对进行 NAT 操作的 IP 分组的净荷没有要求。

8.3.3　实现内部网络之间通信

1. 基本情况

网络结构如图 8.15 所示，两个内部网络通过公共网络互连，由于这两个内部网络相互独立，可以分配相同的私有地址块 192.168.1.0/24，但在建立私有地址与全球地址之间映射前，其他网络中的终端无法用某个终端的私有地址访问该终端，因此，必须由内部网络中分配私有地址的终端发起访问公共网络中分配全球 IP 地址的终端的过程，如果需要实现内部网络 1 中配置私有 IP 地址 192.168.1.1 的终端访问内部网络 2 中配置私有 IP 地址 192.168.1.1的服务器，需要在路由器 R2 建立私有 IP 地址 192.168.1.1 和全球 IP 地址 193.1.3.1 之间的静态映射。路由器 R1 中必须建立用于指明通往目的网络 193.1.3.0/28 的传输路径的路由项。路由器 R2 中必须建立用于指明通往目的网络 193.1.1.16/28 的传输路径的路由项。当内部网络 1 中的终端发起访问内部网络 2 中服务器时，构建并发送以私有地址 192.168.1.1 为源 IP 地址、以全球地址 193.1.3.1 为目的 IP 地址的 IP 分组，需由路由器 R1 完成该 IP 分组源 IP 地址私有地址至全球地址的转换，并建立地址转换项。由路由器 R2 根据私有地址 192.168.1.1 和全球地址 193.1.3.1 之间的静态映射将该 IP 分组的目的 IP 地址转换成私有地址 192.168.1.1。

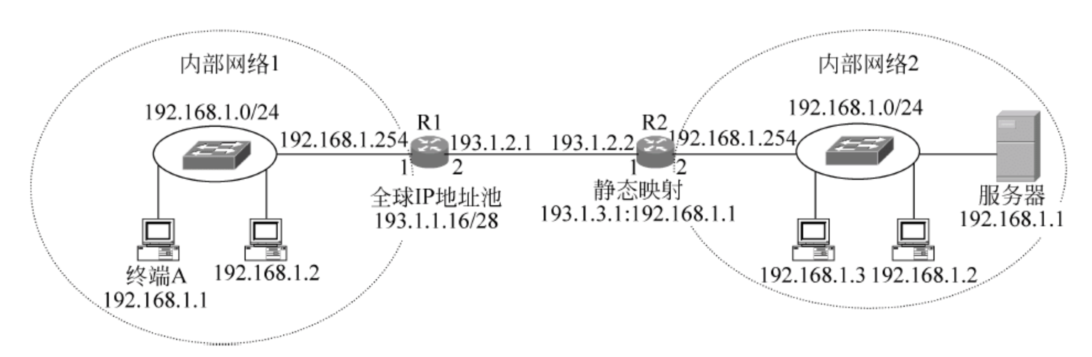


图 8.15　网络结构

2. 基本配置

1) 路由器路由表(表 8.7、表 8.8)

对于内部网络 1 中的终端，内部网络 2 中的服务器的 IP 地址属于全球地址 193.1.3.0/28。对于内部网络 2 中的服务器，内部网络 1 中的终端的 IP 地址属于全球地址 193.1.1.16/28。内部网络 1 私有地址与全球地址之间映射由动态 NAT 实现，内部网络 2 私有地址与全球地址之间映射由静态 NAT 实现。

表 8.7　路由器 R1 路由表

目 的 网 络	下一跳
192.168.1.0/24	直接
193.1.3.0/28	193.1.2.2

表 8.8 路由器 R2 路由表

目 的 网 络	下一跳
192.168.1.0/24	直接
193.1.1.16/28	193.1.2.1

2）路由器 R1 动态 NAT 配置

①配置全球 IP 地址池 193.1.1.16/28。②配置路由器连接内部网络和外部网络接口。③指定使用动态 NAT，并配置允许进行地址转换的内部网络本地地址（私有地址）范围。

3）路由器 R2 静态 NAT 配置

①配置路由器连接内部网络和外部网络接口。②指定使用静态 NAT，并建立私有地址 192.168.1.1 与全球地址 193.1.3.1 之间的静态映射。

3. IP 分组传输过程

下面以内部网络 1 中的终端 A 访问内部网络 2 中的服务器为例，讨论两个内部网络之间的 IP 分组传输过程。内部网络 1 中的终端 A 向内部网络 2 中的服务器传输 IP 分组的过程如下。

- 终端 A 构建以本地地址 192.168.1.1 为源 IP 地址、以全球地址 193.1.3.1 为目的 IP 地址的 IP 分组，值得强调的是，对于内部网络 1 中的终端，内部网络 2 中的服务器的 IP 地址是全球 IP 地址 193.1.3.1。因此，路由器 R2 必须事先建立本地地址 192.168.1.1（内部本地地址）与全球地址 193.1.3.1（内部全球地址）之间的映射。
- 终端 A 通过配置的默认网关地址将该 IP 分组传输给路由器 R1，路由器 R1 确定通过连接内部网络的接口接收到该 IP 分组，用该 IP 分组的目的 IP 地址 193.1.3.1 检索路由表，找到匹配的路由项，发现该路由项中的输出接口是连接外部网络的接口，并且该 IP 分组的源 IP 地址属于允许进行 NAT 操作的本地地址范围（192.168.1.1∈192.168.1.0/24），对该 IP 分组实施 NAT 操作，在全球地址 IP 地址池中选择一个未分配的全球 IP 地址（这里假定是 193.1.1.17）作为该 IP 分组的源 IP 地址，在地址转换表中创建一项用于建立内部本地地址与内部全球地址之间映射的地址转换项 192.168.1.1（内部本地地址）：193.1.1.17（内部全球地址）。
- 路由器 R1 根据匹配的路由项，将该 IP 分组传输给路由器 R2，由于路由器 R2 确定通过连接外部网络的接口接收到该 IP 分组，在地址转换表中检索内部全球地址与该 IP 分组的目的 IP 地址相同的地址转换项，并用该地址转换项中的内部本地地址作为该 IP 分组的目的 IP 地址。由于路由器 R2 的地址转换表中存在静态地址转换项 193.1.3.1（内部全球地址）：192.168.1.1（内部本地地址），该 IP 分组的目的 IP 地址转换为 192.168.1.1。路由器 R2 用该目的 IP 地址检索路由表，找到匹配的路由项，根据该路由项，将该 IP 分组传输给服务器。值得强调的是，该 IP 分组到达服务器时，源 IP 地址是全球 IP 地址 193.1.1.17，目的 IP 地址是本地地址 192.168.1.1。即目的 IP 地址是服务器的本地地址，源 IP 地址是内部网络 2 标识终端 A 的全球 IP 地址。

内部网络 2 中的服务器向内部网络 1 中的终端 A 传输 IP 分组的过程如下。

- 服务器构建以本地地址 192.168.1.1 为源 IP 地址、以全球地址 193.1.1.17 为目的 IP 地址的 IP 分组，这意味着内部网络 2 用全球 IP 地址 193.1.1.17 标识终端 A，这是路由器 R1 已经建立内部本地地址 192.168.1.1 与内部全球地址 193.1.1.17 之间映射为前提的。
- 服务器通过配置的默认网关地址将该 IP 分组传输给路由器 R2，路由器 R2 确定通过连接内部网络的接口接收到该 IP 分组，用该 IP 分组的目的 IP 地址 193.1.1.17 检索路由表，找到匹配的路由项，发现该路由项中的输出接口是连接外部网络的接口，并且该 IP 分组的源 IP 地址属于允许进行 NAT 操作的本地地址范围(192.168.1.1∈192.168.1.0/24)，对该 IP 分组实施 NAT 操作。由于地址转换表中已经存在地址转换项 192.168.1.1(内部本地地址)：193.1.3.1(内部全球地址)，路由器 R2 用内部全球 IP 地址 193.1.3.1 作为该 IP 分组的源 IP 地址。
- 路由器 R2 根据匹配的路由项，将该 IP 分组传输给路由器 R1，由于路由器 R1 确定通过连接外部网络的接口接收到该 IP 分组，在地址转换表中检索内部全球地址与该 IP 分组的目的 IP 地址相同的地址转换项，并用该地址转换项中的内部本地地址作为该 IP 分组的目的 IP 地址。由于路由器 R1 的地址转换表中已经存在地址转换项 192.168.1.1(内部本地地址)：193.1.1.17(内部全球地址)，该 IP 分组的目的 IP 地址转换为 192.168.1.1。路由器 R1 用该目的 IP 地址检索路由表，找到匹配的路由项，根据该路由项，将该 IP 分组传输给终端 A。

8.3.4 解决内部网络与外部网络地址重叠问题

1. 基本情况

内部网络与外部网络地址重叠情况如图 8.16 所示。内部网络包含子网 192.168.1.0/24 和 192.168.2.0/24，外部网络包含子网192.168.2.0/24，如果内部网络中的终端 A 直接以 IP 地址 192.168.2.1 访问外部网络中的服务器，终端 A 发送的以 192.168.1.1 为源 IP 地址、以 192.168.2.1 为目的 IP 地址的 IP 分组被路由器 R1 直接转发给终端 B，无法到达外部网络中的 Web 服务器。同样，内部网络中的私有地址空间(192.168.1.0/24 和 192.168.2.0/24)对路由器 R2 是透明的，路由器 R2 只能将目的 IP 地址属于全球 IP 地址 193.1.1.16/28 的 IP 分组转发给路由器 R1。这种情况下，路由器 R1 需要定义全球 IP 地址池 193.1.1.16/28，建立外部网络 Web 服务器外部全球地址 192.168.2.1 与外部本地地址 193.1.3.1 之间的静态映射。这里的外部全球地址指的是 Web 服务器在外部网络使用的 IP 地址，外部本地地址指的是 Web 服务器在内部网络使用的 IP 地址。当终端 A 发起访问外部网络中的 Web 服务器时，终端 A 构建并发送以 192.168.1.1 为源 IP 地址、以 193.1.3.1为目的 IP 地址的 IP 分组，当该 IP 分组到达路由器 R1 时，如果满足以下全部条件：

- 路由器 R1 通过连接内部网络的接口接收到该 IP 分组；
- 通过检索路由表确定该 IP 分组通过路由器连接外部网络的接口输出；
- 地址转换表中存在外部本地地址等于 193.1.3.1 的地址转换项；
- IP 分组源 IP 地址属于允许进行地址转换的内部网络本地地址范围；
- 已经定义了用于转换内部网络本地地址的全球 IP 地址池。

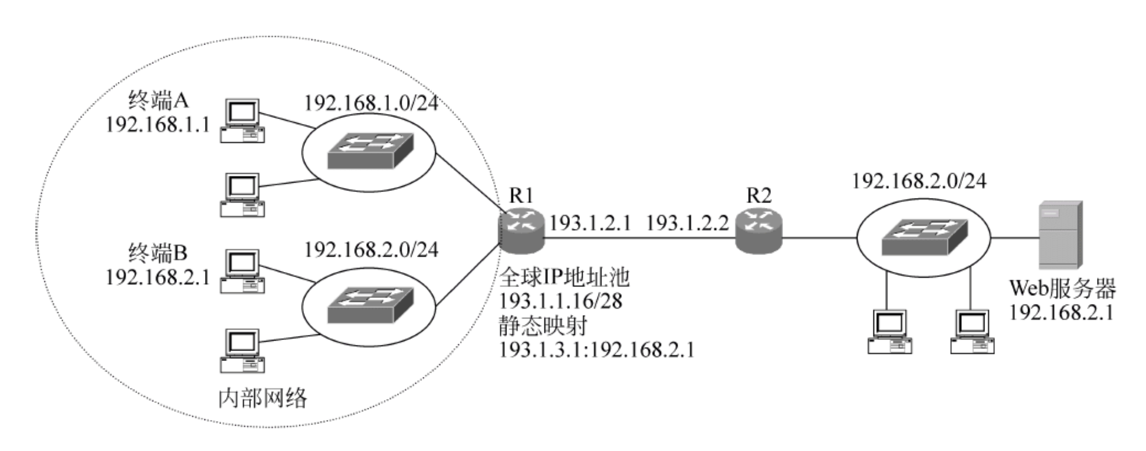


图 8.16 内部网络与外部网络地址重叠情况

路由器 R1 在全球地址池中选择一个全球 IP 地址(这里假定是 193.1.1.17),用该全球 IP 地址取代 IP 分组中的源 IP 地址,用外部本地地址等于 193.1.3.1 的地址转换项中给出的外部全球地址 192.168.2.1 取代该 IP 分组的目的 IP 地址,将完成地址转换后的 IP 分组发送给路由器 R2。同时,在地址转换表中增加内部本地地址为 192.168.1.1、内部全球地址为 193.1.1.17、外部本地地址为 193.1.3.1、外部全球地址为 192.168.2.1 的地址转换项。

Web 服务器发送给终端 A 的 IP 分组的源 IP 地址是 192.168.2.1、目的 IP 地址是 193.1.1.17,当路由器 R1 通过连接外部网络的接口接收到该 IP 分组,在地址转换表中检索内部全球地址等于该 IP 分组目的 IP 地址、外部全球地址等于该 IP 分组源 IP 地址的地址转换项,用该地址转换项的内部本地地址取代该 IP 分组的目的 IP 地址、用该地址转换项的外部本地地址取代该 IP 分组的源 IP 地址,然后对完成地址转换后的 IP 分组进行转发操作。

建立 Web 服务器外部全球地址 192.168.2.1 与外部本地地址 193.1.3.1 之间的静态映射,和终端 A 向 Web 服务器发送以 193.1.3.1 为目的 IP 地址的 IP 分组后,路由器 R1 的地址转换表如表 8.9 所示。

表 8.9 路由器 R1 地址转换表

内部本地地址	内部全球地址	外部本地地址	外部全球地址
		193.1.3.1	192.168.2.1
192.168.1.1	193.1.1.17	193.1.3.1	192.168.2.1

2. 基本配置

1) 路由器路由表(表 8.10、表 8.11)

对于路由器 R2,内部网络的地址范围是路由器 R1 全球 IP 地址池的 IP 地址范围,因此所有传输给内部网络的 IP 分组的目的 IP 地址属于全球 IP 地址 193.1.1.16/28。同样,对于内部网络,发送给外部网络中属于子网 192.168.2.0/24 的终端或服务器的 IP 分组的目的 IP 地址属于 193.1.3.0/24,因此路由器 R1 在地址转换前需要把以属于 193.1.3.0/24 的 IP 地址为目的 IP 地址的 IP 分组转发给路由器 R2。

表 8.10　路由器 R1 路由表

目 的 网 络	下一跳
192.168.1.0/24	直接
192.168.2.0/24	直接
193.1.3.0/24	193.1.2.2

表 8.11　路由器 R2 路由表

目 的 网 络	下一跳
192.168.2.0/24	直接
193.1.1.16/28	193.1.2.1

2）路由器 R1 动态 NAT 配置

路由器 R1 动态 NAT 配置如下。

①配置全球 IP 地址池 193.1.1.16/28。②配置路由器连接内部网络和外部网络接口。③指定使用动态 NAT,并配置允许进行地址转换的内部网络本地地址(私有地址)范围。

3）路由器 R1 静态 NAT 配置

路由器 R1 静态 NAT 配置如下。

指定使用静态 NAT,并建立外部全球地址 192.168.2.1 与外部本地地址 193.1.3.1 之间的静态映射。

3. IP 分组传输过程

下面以内部网络中的终端 A 访问外部网络中的 Web 服务器为例讨论内部网络与外部网络之间的 IP 分组传输过程。内部网络中的终端 A 向外部网络中的 Web 服务器传输 IP 分组过程如下。

- 终端 A 构建以本地地址 192.168.1.1 为源 IP 地址、以外部本地地址 193.1.3.1 为目的 IP 地址的 IP 分组,193.1.3.1 是内部网络用于标识外部网络中的 Web 服务器的 IP 地址,因而被称为外部本地地址。因此,路由器 R1 必须事先建立外部本地地址 193.1.3.1 与外部全球地址 192.168.1.1 之间的映射。
- 当路由器 R1 通过连接内部网络的接口接收到该 IP 分组,用目的 IP 地址 193.1.3.1 检索路由表,发现匹配的路由项的输出接口是连接外部网络的接口,且该 IP 分组的源 IP 地址属于允许进行 NAT 操作的本地地址范围,路由器 R1 在全球 IP 地址池中选择一个未分配的全球 IP 地址(这里假定是 193.1.1.17),用该全球 IP 地址作为该 IP 分组的源 IP 地址,创建用于建立内部本地地址 192.168.1.1 与内部全球地址 193.1.1.17 之间映射的地址转换项。如果该 IP 分组的目的 IP 地址等于某项地址转换项中的外部本地地址,用该地址转换项中的外部全球地址作为该 IP 分组的目的 IP 地址。由于路由器 R1 中已经存在用于建立外部本地地址 193.1.3.1 与外部全球地址 192.168.2.1 之间映射的地址转换项,完成 NAT 操作后的 IP 分组的源 IP 地址为 193.1.1.17、目的 IP 地址为 192.168.2.1。路由器 R1 通过连接外部网络的接口输出该 IP 分组。

外部网络中的 Web 服务器向内部网络中的终端 A 传输 IP 分组过程如下。

- Web 服务器构建以 192.168.2.1 为源 IP 地址、以 193.1.1.17 为目的 IP 地址的 IP 分组。
- 当路由器 R1 通过连接外部网络的接口接收到该 IP 分组，如果在地址转换表中检索到内部全球地址等于该 IP 分组的目的 IP 地址的地址转换项，用该地址转换项的内部本地地址作为该 IP 分组的目的 IP 地址。如果在地址转换表中检索到外部全球地址等于该 IP 分组的源 IP 地址的地址转换项，用该地址转换项的外部本地地址作为该 IP 分组的源 IP 地址。完成 NAT 操作后的该 IP 分组的源 IP 地址为 193.1.3.1、目的 IP 地址为 192.168.1.1。路由器 R1 用目的 IP 地址 192.168.1.1 检索路由表，根据匹配的路由项，将 IP 分组传输给终端 A。

值得再次强调的是，如果边界路由器通过连接内部网络的接口接收到 IP 分组，其操作步骤是首先检索路由表、确定输出接口，在确定输出接口是连接外部网络的接口后，完成地址转换操作，检索路由表的操作在地址转换操作之前。如果边界路由器通过连接外部网络的接口接收到 IP 分组，首先完成地址转换操作，然后检索路由表，检索路由表的操作在地址转换操作之后。

习题

8.1 NAT 能够缓解 IP 地址短缺问题的原因是什么？

8.2 NAT 对提高网络安全有什么帮助？

8.3 NAT 对网络通信有什么副作用？如何解决？

8.4 NAT 和 PAT 有什么本质区别？各自适用什么网络环境？

8.5 实现应用层网关的困难是什么？

8.6 不同的内部网络能否采用相同的本地 IP 地址？如果两个内部网络分配了相同的本地 IP 地址，会对两个内部网络中的终端之间的通信过程带来麻烦吗？

8.7 图 8.17 如何配置路由器 R1、路由器 R2 的 NAT，才能实现终端 A 和终端 C 之间的相互通信。

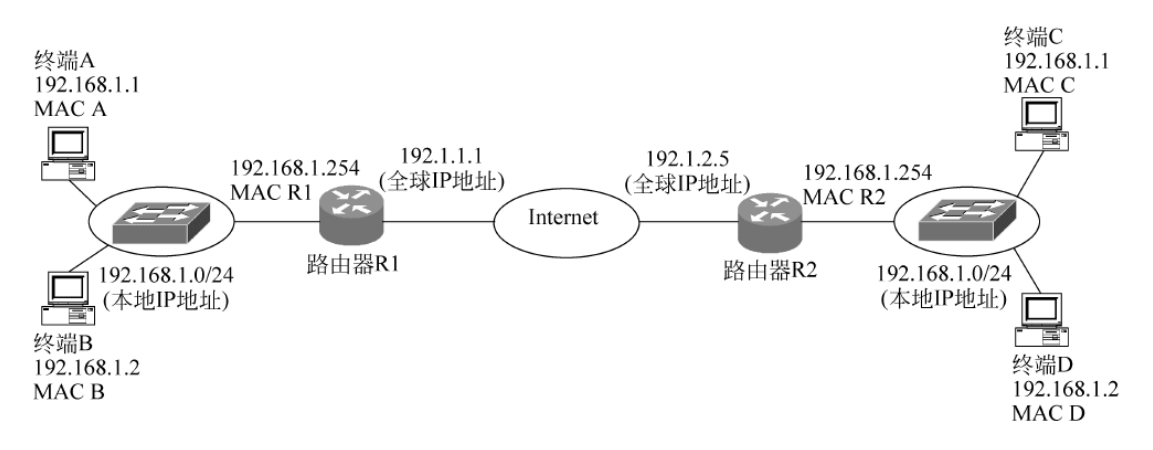


图 8.17 题 8.7 图

8.8 对应图 8.18 所示的网络结构和 IP 地址配置，给出能够实现终端 A 与 Web 服务器

2、终端 B 与 Web 服务器 1 之间通信的配置(包括路由器路由表和路由器 R1 的 NAT 配置)。

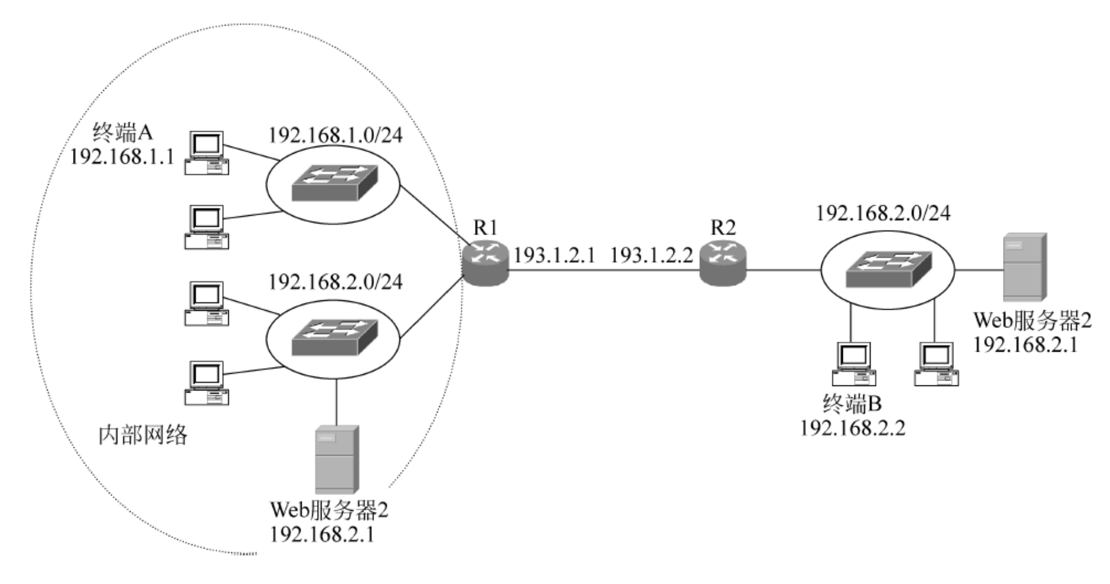


图 8.18　题 8.8 图

8.9　对应图 8.19 所示的网络结构和 IP 地址配置,给出能够实现内部网络终端访问 Web 服务器 2、外部网络终端访问 Web 服务器 1 所需要的配置(包括路由器路由表和路由器 R1 的 NAT 配置)。

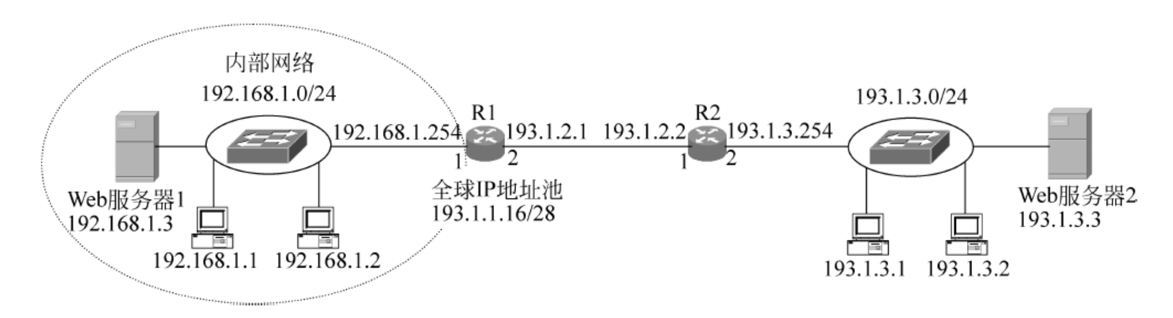


图 8.19　题 8.9 图



第9章 三层交换机和三层交换

为了解决大型交换式以太网引发的广播风暴和安全问题，将单个物理交换式以太网划分为多个虚拟局域网（VLAN），这些VLAN虽然共享同一个物理交换式以太网，但逻辑上是相互独立的，需要通过网络层互连设备实现VLAN间通信，当然可以用路由器互连VLAN，但基于VLAN的特殊性，路由器并不是互连VLAN的最佳设备，实现交换式以太网VLAN划分和VLAN间通信的最佳设备是集二层交换和三层路由于一体的三层交换机。

9.1 三层交换机基础

9.1.1 三层交换机产生背景

1. 路由器用不同的物理接口连接不同的VLAN

在第2章讨论VLAN时已经讲到，为了缩小广播域，通过划分VLAN的方法将一个交换式以太网划分成多个相互隔离的广播域，但每一个广播域是逻辑上独立的网络，需要通过网络层互连设备实现这些广播域之间的通信，路由器是传统的网络层互连设备。图9.1就是一个通过路由器实现三个VLAN之间相互通信的互连网络结构。

一般情况下，路由器的每一个接口需要分配IP地址和子网掩码，接口分配的IP地址和子网掩码决定了该接口连接的网络的网络地址。同时，每一个路由器接口还具有一个和该接口连接的网络相关的物理地址，对于如图9.1所示的路由器每一个接口连接以太网的情况，路由器每一个接口还具有一个MAC地址。图9.1中，用一个带三个以太网接口的路由器来互连三个VLAN，为路由器三个以太网接口分配的IP地址和子网掩码决定了接口连接的VLAN的网络地址。当VLAN 2中IP地址为192.1.1.1的终端需要向VLAN 4中IP地址为192.1.3.1的终端发送IP分组时，执行下述操作过程：

① 通过和子网掩码进行“与”操作，确定源和目的终端不在同一个子网，源终端首先将IP分组发送给默认网关：192.1.1.254。IP地址192.1.1.254是路由器连接VLAN 2的接口的IP地址。由于源终端和该接口之间的传输网络是以太网（VLAN 2），因此，源终端通过ARP地址解析过程，获取该接口的MAC地址MAC R1，并因此构建以MAC 1为源MAC地址，MAC R1为目的MAC地址的MAC帧，通过以太网将该MAC帧传输给默认网关。

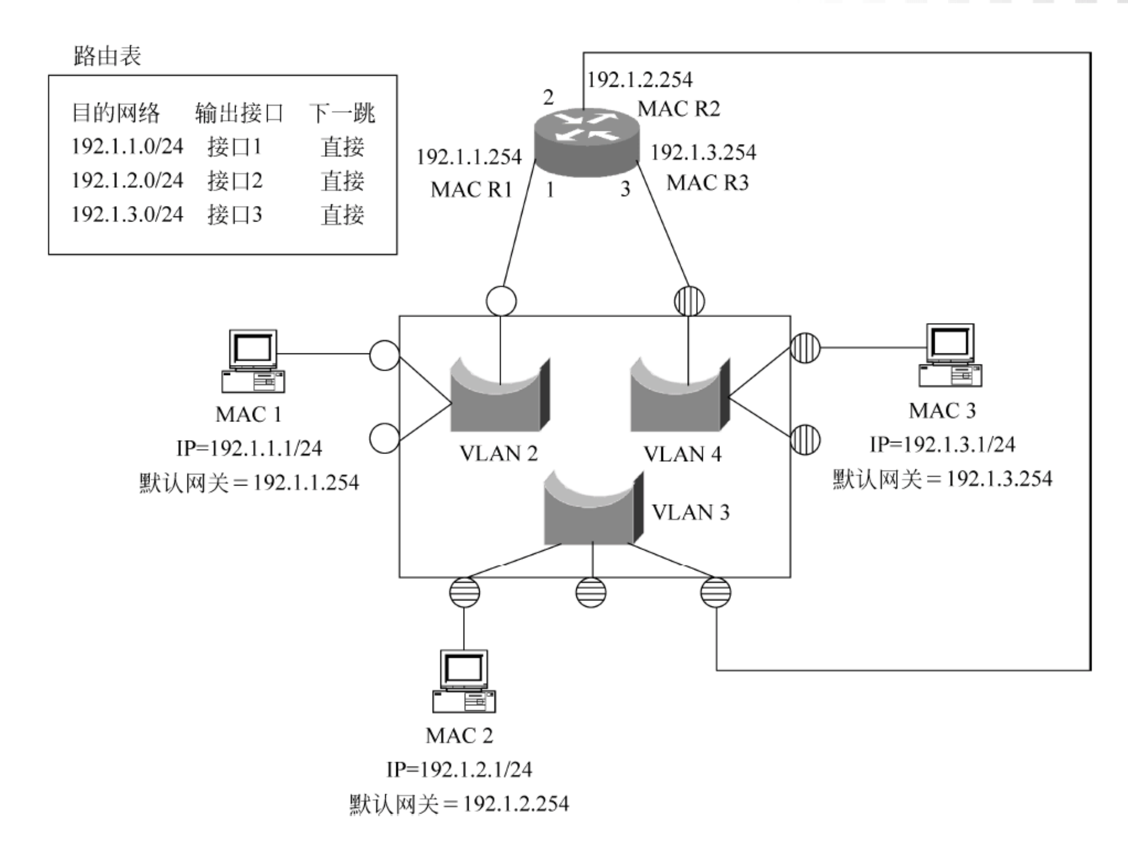


图 9.1　用路由器实现三个 VLAN 之间通信

② 默认网关从 MAC 帧中分离出 IP 分组，根据 IP 分组的目的 IP 地址去检索路由表，找到匹配的路由项＜192.1.3.0/24，接口 3，直接＞。通过在接口 3 连接的以太网进行的 ARP 地址解析过程，直接获取目的的终端（IP 地址＝192.1.3.1）的 MAC 地址，在获取目的终端的 MAC 地址（MAC 3）后，构建以默认网关连接目的终端所在网络的接口的 MAC 地址（MAC R3）为源 MAC 地址，目的终端 MAC 地址（MAC 3）为目的 MAC 地址的 MAC 帧，并通过以太网将该 MAC 帧传输给目的终端。

图 9.1 所示的互连网络结构直观、简单、容易理解，但不易在具体的网络设计中实现，一是由于属于每一个 VLAN 的交换机端口在交换式以太网中是任意分布的，而路由器的每一个物理接口又必须连接对应 VLAN 的非标记端口（接入端口），导致路由器接口与对应 VLAN 之间的物理连接难以实现。二是由于 VLAN 的划分是动态变化的，因此，无法在设计、实施网络时确定路由器的以太网接口数。三是由于需要增加路由器来实现 VLAN 之间通信，实现 VLAN 之间通信的成本较高。

2. 路由器用单个物理接口连接不同的 VLAN

为了解决用不同的路由器物理接口连接不同的 VLAN 所带来的问题，将图 9.1 所示的互连网络结构转换成图 9.2 所示的互连网络结构。

图 9.2 所示的互连网络结构中，路由器连接以太网的物理接口被划分成 3 个逻辑接口，每个逻辑接口连接一个 VLAN，因此，这 3 个逻辑接口分别连接 VLAN 2、VLAN 3、VLAN 4，分别有了 VLAN 标识符 2、3、4，也分别配置了和 VLAN 2、VLAN 3、VLAN 4 网络地址

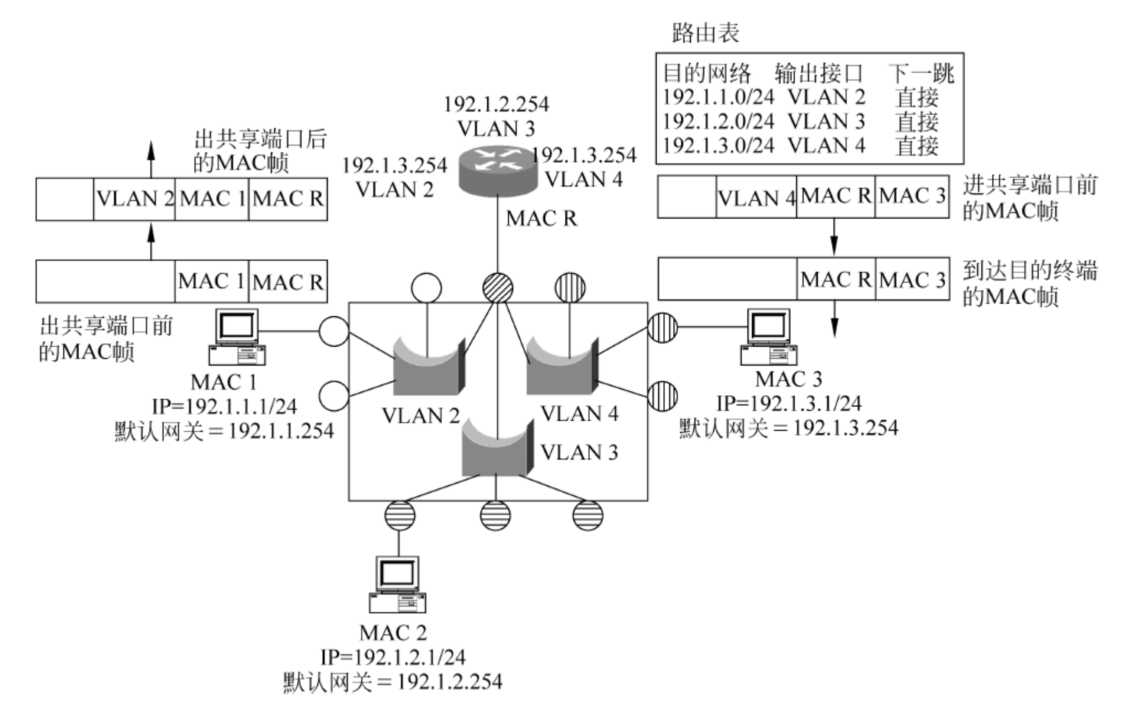


图 9.2 路由器单一物理接口划分成多个逻辑接口的方法

一致的接口地址：192.1.1.254(VLAN 2 网络地址＝191.1.1.0/24)、192.1.2.254(VLAN 3 网络地址＝192.1.2.0/24)和 192.1.3.254(VLAN 4 网络地址＝192.1.3.0/24)。逻辑接口是功能上完全等同于一个独立的物理接口，但物理上必须和其他逻辑接口共享一个物理接口的一种连接外部网络的方式。物理接口通过接收到的 MAC 帧所携带的 VLAN 标识符来确定真正接收该 MAC 帧的逻辑接口，因此，发送给路由器的 MAC 帧必须携带 VLAN 标识符。以太网交换机中连接路由器的以太网端口必须是被三个 VLAN 共享的共享端口，且是 802.1Q 标记端口。当图 9.2 中 VLAN 2 中 IP 地址为 192.1.1.1 的终端希望向 VLAN 4 中 IP 地址为 192.1.3.1 的终端发送 IP 分组时，执行以下操作过程。

① 通过和子网掩码进行“与”操作，确定源和目的终端不在同一网络，源终端首先将 IP 分组发送给默认网关。为了获取默认网关的 MAC 地址，源终端在 VLAN 2 内广播 ARP 请求帧，该 MAC 帧通过所有属于 VLAN 2 的端口发送出去，包括被 3 个 VLAN 共享的以太网交换机端口，通过该共享端口发送出去的 MAC 帧携带 VLAN 标识符——VLAN 2。路由器回送默认网关的 MAC 地址时，也使 ARP 响应帧携带 VLAN 标识符——VLAN 2。该 MAC 帧进入被 3 个 VLAN 共享的以太网交换机共享端口时，以太网交换机通过该 MAC 帧携带的 VLAN 标识符获知用于转发该 MAC 帧的网桥，通过和 VLAN 2 相关联的网桥将该 MAC 帧转发给源终端。源终端构建以自身 MAC 地址(MAC 1)为源 MAC 地址，默认网关 MAC 地址(MAC R)为目的 MAC 地址的 MAC 帧，并将该 MAC 帧传输给以太网，同样，当该 MAC 帧从被 3 个 VLAN 共享的以太网交换机共享端口转发出去时，携带 VLAN 标识符——VLAN 2。

② 路由器接收到该 MAC 帧，从中分离出 IP 分组，根据 IP 分组的目的 IP 地址去检索路由表，找到匹配的路由项＜192.1.3.0/24，VLAN 4，直接＞。为了获取目的终端的 MAC

地址，路由器也构建 ARP 请求帧，并使得该 ARP 请求帧携带 VLAN 标识符——VLAN 4。当路由器发送的 ARP 请求帧进入被 3 个 VLAN 共享的以太网交换机共享端口时，以太网交换机通过其携带的 VLAN 标识符获悉它所属的 VLAN，因此，该 MAC 帧只在 VLAN 4 中广播。目的终端将自身的 MAC 地址通过 ARP 响应帧回送给路由器。路由器构建一个以自身 MAC 地址(MAC R)为源 MAC 地址，目的终端 MAC 地址(MAC 3)为目的 MAC 地址的 MAC 帧，为该 MAC 帧加上 VLAN 标识符——VLAN 4，并将该 MAC 帧发送给以太网。当该 MAC 帧通过被 3 个 VLAN 共享的以太网交换机共享端口进入以太网交换机时，以太网交换机通过其携带的 VLAN 标识符找到和该 VLAN 相关联的网桥，并由该网桥将该 MAC 帧转发给目的终端。

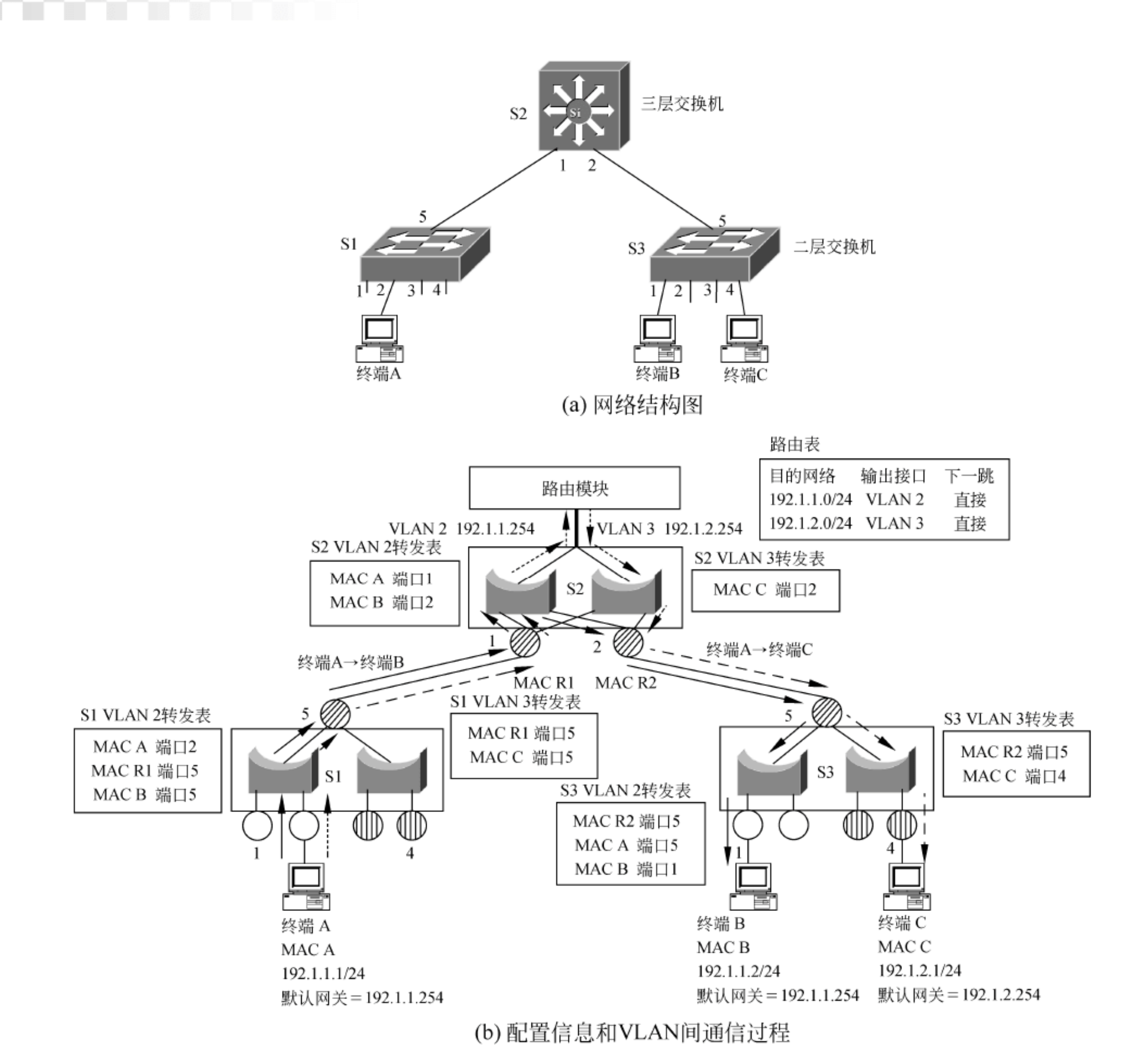
图 9.2 所示的互连网络结构称为单臂路由器结构，只需将路由器物理接口与交换式以太网中某个被所有 VLAN 共享且是 802.1Q 标记端口的交换机端口相连，就可实现所有 VLAN 之间的通信。这种互连网络结构解决了用不同的路由器物理接口连接不同的 VLAN 所带来的路由器物理接口必须连接对应 VLAN 的接入端口和路由器物理接口数目随 VLAN 变化而变化的问题。但仍然存在如下问题，一是仍然需要通过路由器实现 VLAN 之间通信，实现 VLAN 之间通信的成本较高。二是由于所有 VLAN 间流量都需经过路由器物理接口与交换式以太网共享端口之间的物理链路，使得该物理链路成为性能瓶颈，尤其当 VLAN 间流量占网络总流量的比例较高时，该问题更加突出。

3. 三层交换机实现 VLAN 间通信过程

目前功能强一点的以太网交换机都采用机箱式结构，机箱内装有背板，各个功能模块插在背板上，通过背板实现功能模块之间通信，背板的带宽设计得非常高。这种情况下，以太网交换机厂商自然想到通过在以太网交换机中增加一个路由模块，将以太网交换机变成一个集交换、路由功能于一体的新设备——三层交换机。将一个集交换、路由功能于一体的新设备称作三层交换机的原因是路由功能是网络层的功能，而在基于以太网的 TCP/IP 体系结构中，网络层位于第三层，因此，将具有路由功能的设备称作三层设备，而将只有 MAC 层功能的设备称作二层设备，也有了二层交换机和三层交换机的叫法。当然，目前情况下，并不是只有机箱式以太网交换机才有可能是三层交换机，许多固定端口的以太网交换机也安装了路由模块。用三层交换机实现 VLAN 间通信的互连网络结构如图 9.3(a)所示，对应的配置信息和 VLAN 间通信过程如图 9.3(b)所示。

图 9.3 中的三层交换机主要由两部分组成：支持 VLAN 划分的二层交换结构和路由模块，两者之间通过背板完成信息交换。路由模块的功能就像一个传统的路由器，运行路由协议，建立路由表，完成 IP 分组转发等。而二层交换结构就像普通以太网交换机一样，用目的 MAC 地址检索转发表，根据转发表给出的路由信息转发 MAC 帧。

假定图 9.3(a)中的各个交换机已经建立了如图 9.3(b)所示的转发表，同一 VLAN 内两个终端之间的通信就像在普通交换式以太网中通信一样，不需要涉及路由模块。如终端 A→终端 B 之间的通信，终端 A 将以 MAC A 为源 MAC 地址，MAC B 为目的 MAC 地址的 MAC 帧发送给以太网交换机 1(S1)，以太网交换机 1 根据接收该 MAC 帧的端口确定该 MAC 帧属于 VLAN 2，由和 VLAN 2 关联的网桥转发该 MAC 帧。和 VLAN 2 关联的网桥检索对应的转发表，找到转发端口(端口 5)，由于转发端口被两个 VLAN 所共享且被配



(a) 网络结构图

(b) 配置信息和VLAN间通信过程

图 9.3 用三层交换机实现 VLAN 间通信

置为 802.1Q 标记端口，因此，将该 MAC 帧从端口 5 转发出去之前，先加上 VLAN 标识符——VLAN 2。从以太网交换机 1 端口 5 转发出去的 MAC 帧通过端口 1 进入以太网交换机 2(S2)。以太网交换机 2 通过该 MAC 帧携带的 VLAN 标识符——VLAN 2 确定该 MAC 帧所属的 VLAN，并将该 MAC 帧提交给和 VLAN 2 关联的网桥进行转发。和 VLAN 2 关联的网桥通过检索对应的转发表，找到转发端口（端口 2），由于转发端口也是一个被两个 VLAN 所共享且被配置为 802.1Q 标记端口的共享端口，因此，从该端口转发出去的 MAC 帧仍然携带 VLAN 标识符——VLAN 2。同样，该 MAC 帧进入以太网交换机 3(S3)后，确定由和 VLAN 2 关联的网桥转发该 MAC 帧，并通过检索 VLAN 2 对应的转发表找到转发端口，由于转发端口（端口 1）是一个非标记端口，从这样的端口转发出去的 MAC 帧必须去除 VLAN 标识符。没有携带 VLAN 标识符的 MAC 帧通过端口 1 到达终端 B，完成了 MAC 帧终端 A→终端 B 的传输过程。

为了实现属于不同 VLAN 的终端之间的通信，必须对路由模块进行配置：建立两个逻辑接口，分别对应 VLAN 2 和 VLAN 3，为这两个逻辑接口分配和对应 VLAN 网络地址一致的接口地址，如图 9.3(b)中，逻辑接口 1 对应 VLAN 2，分配接口地址 192.1.1.254

(VLAN 2 的网络地址为 192.1.1.0/24),逻辑接口 2 对应 VLAN 3,分配接口地址 192.1.2.254(VLAN 3 的网络地址为 192.1.2.0/24)。完成上述配置后,路由模块将自动在路由表中增添两项路由项,如图 9.3(b)所示。这种情况下,通过三层交换机可以实现属于不同 VLAN 的两个终端之间的通信。下面以终端 A→终端 C 通信过程为例,讨论一下三层交换机实现 VLAN 之间通信的过程。

① 终端 A 通过将自身的 IP 地址和目的终端(终端 C)的 IP 地址与子网掩码进行"与"操作后发现,源终端和目的终端不在同一个子网,终端 A 确定需要将 IP 分组先转发给默认网关。为了获取默认网关的 MAC 地址,终端 A 广播一个 ARP 请求帧,该 ARP 请求帧到达 VLAN 2 内所有终端和路由模块。路由模块发现 ARP 请求帧中要求解析的 IP 地址是自己的接口地址,但路由模块本身没有端口,因此也不会有 MAC 地址,由于三层交换机中的每一个端口均有 MAC 地址,路由模块将三层交换机接收该 ARP 请求帧的端口的 MAC 地址(MAC R1)作为自身的 MAC 地址,回复给终端 A。终端 A 就构建一个以自身 MAC 地址(MAC A)为源 MAC 地址,以太网交换机 2 端口 1 的 MAC 地址(MAC R1)为目的 MAC 地址的 MAC 帧,并将该 MAC 帧发送给以太网。该 MAC 帧最终进入以太网交换机 2 端口 1,由于该 MAC 帧的目的 MAC 地址是以太网交换机 2 自身端口的 MAC 地址,以太网交换机 2 直接将这样的 MAC 帧转发给路由模块,而不是通过和某个 VLAN 关联的网桥进行转发操作。

② 路由模块从该 MAC 帧中分离出 IP 分组,用该 IP 分组的目的 IP 地址检索路由表,获知可以直接通过以太网将 IP 分组转发给目的终端。也通过广播 ARP 请求帧来获取目的终端的 MAC 地址,该 ARP 请求帧在 VLAN 3 中广播,到达 VLAN 3 中的所有终端。但从不同的以太网交换机 2 端口中发送出来的 ARP 请求帧,其源 MAC 地址是不一样的,因为所有由路由模块发送的 MAC 帧,都用发送该 MAC 帧的以太网交换机端口的 MAC 地址作为该 MAC 帧的源 MAC 地址。终端 C 接收到该 ARP 请求帧后,回复一个响应帧,并将其 MAC 地址告知路由模块。路由模块构建一个以终端 C 的 MAC 地址(MAC C)为目的 MAC 地址,并携带 VLAN 标识符——VLAN 3 的 MAC 帧,并将该 MAC 帧提交给和 VLAN 3 关联的网桥。接下来以二层交换的方式将该 MAC 帧转发给终端 C,完成了不同 VLAN 之间的通信过程。

三层交换机与路由器实现 VLAN 间通信的主要区别如下:一是所有属于某个 VLAN 的交换机端口与路由模块所在的交换机之间存在交换路径,该交换路径完全等同于属于相同 VLAN 的两个交换机端口之间的交换路径;二是路由模块所在的交换机通过背板实现二层交换路径与路由模块之间的通信,背板的带宽远高于互连路由器接口和交换机端口之间的物理链路的带宽;三是交换机因为增加路由模块而增加的成本远低于因为增加路由器而增加的成本;四是鉴于 VLAN 之间通信的特殊性,三层交换机能够比路由器更快速地完成不同 VLAN 之间的 MAC 帧转发过程。鉴于上述原因,三层交换机已成为实现 VLAN 之间通信的主流产品。

9.1.2　三层交换机与路由器的区别

1. 三层路由与二层交换的有机集成

由于三层交换机集二层交换和三层路由功能于一身,因此允许存在跨三层交换机的

VLAN,对于图 9.4(a)所示的 VLAN 划分,对于两个属于同一 VLAN 的终端之间的通信过程,三层交换机完全等同于二层交换机,对于两个属于不同 VLAN 的终端之间的通信过程,三层交换机实现路由功能。三层交换机根据 MAC 帧的目的 MAC 地址鉴别 MAC 帧的类型,如果该 MAC 帧以三层交换机某个物理端口的 MAC 地址为目的 MAC 地址,该 MAC 帧被直接转发给路由模块,否则,以二层交换方式转发该 MAC 帧。对于路由器(这里讨论的是传统路由器,而不是路由交换机或是交换路由器),每一个物理接口需要连接不同的网络,因此,不可能存在跨路由器的 VLAN,图 9.4(b)中的终端 A 和终端 B 只能属于不同的网络,路由器以路由方式实现终端 A 与终端 B 之间通信。从中可以看出,用路由器分割子网,连接在同一子网的终端之间存在物理地域相关性,这一点与属于同一 VLAN 的交换机端口可以是分布在大型交换式以太网中的任意交换机端口是相悖的。这也是类似校园网这样的大型交换式以太网通过三层交换机,而不是路由器实现 VLAN 分割和 VLAN 间通信的主要原因。

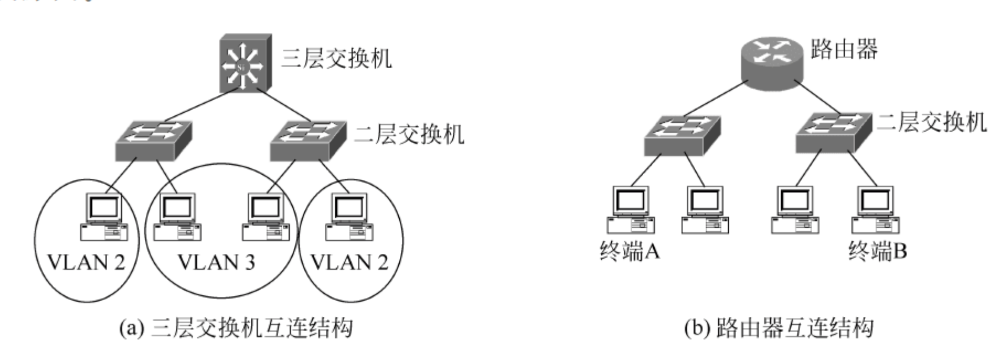


图 9.4 三层交换机与路由器的区别

2. 互连 VLAN 的特殊互连设备

路由器是一种通用的网络层互连设备,可以实现不同网络之间的互连,如以太网和公共交换电话网之间互连,而三层交换机是一种专门用于互连 VLAN 的特殊互连设备。因此,三层交换机主要用于构建大型交换式以太网,实现大型交换式以太网的 VLAN 划分和 VLAN 间通信。

3. IP 接口

三层交换机以某个 VLAN 作为 IP 接口,分配 IP 地址,而一个 VLAN 可以包含多个三层交换机物理端口,发送给某个 IP 接口的 MAC 帧可以从属于对应 VLAN 的任何一个物理端口进入该三层交换机,并由三层交换机转发给该 IP 接口,通过 IP 接口进入路由模块。因此,属于同一 VLAN 的终端与 IP 接口之间交换路径的带宽不受单条物理链路带宽的限制,但连接在某个以太网上的所有终端共享路由器接口连接该以太网的物理链路的带宽。

9.1.3 校园网和三层交换机

1. 校园网特性

校园范围大约在 $2km^2$～$4km^2$ 之间,而且大多数校园是独立、封闭的地理区域,能够实

现自主布线，这两个特点使得交换式以太网成为校园网的最佳组网技术。但校园网中用户种类繁多，有学生、教师、管理者等，信息资源种类繁多，有教学、人事、工资、科研等，用户与信息资源之间存在访问权限分配问题，因此，必须将连接不同类型用户终端的交换机端口划分到不同的 VLAN、将连接不同类型信息资源的交换机端口划分到不同的 VLAN，这样，一是需要实现 VLAN 之间的通信；二是需要对 VLAN 之间的信息交换过程实施控制。三层交换机作为实现 VLAN 间通信的首选设备，自然成为构建校园网的主流设备。

2. 校园网结构

校园网结构如图 9.5 所示，整个网络结构划分为核心层、汇聚层和接入层。核心层由高速主干网组成，其任务是为其他两层提供优化的数据传输功能，由于核心层是分组的总交汇点，必须具有快速交换分组的能力，因此，核心层设备一般不参与可能影响分组交换速率的操作，如分组过滤等。同时，核心层设备也必须尽可能连接高速链路，以免产生带宽瓶颈。

汇聚层提供基于统一策略的互连性，定义网络边界，可以对分组进行复杂的操作，如分组过滤等，同时，它还实现广播域定义、VLAN 间路由等功能。

接入层解决终端设备的接入问题，它一方面需要增大端口密度的技术，如堆叠。另一方面需要对接入进行控制，如 MAC 层过滤、端口安全特性等。

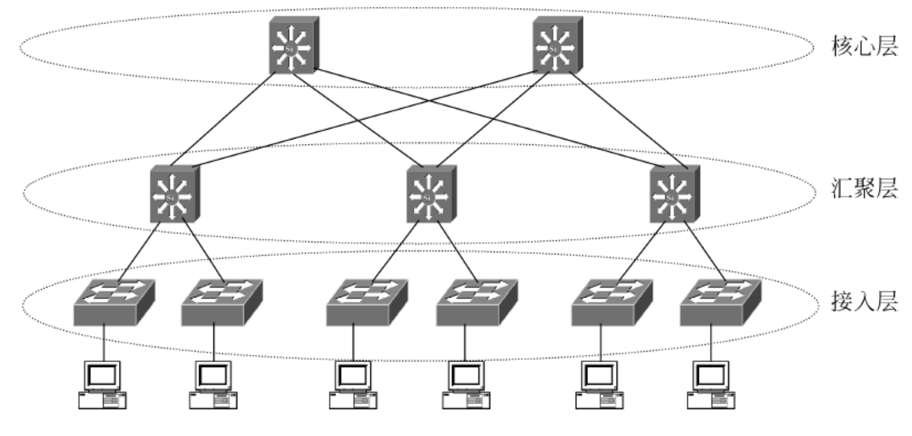


图 9.5　校园网结构

3. 三层交换机在校园网中的应用

图 9.5 中汇聚层和核心层采用三层交换机。每一个汇聚层设备可以连接多个接入层设备，连接在这些接入层设备上的终端可以划分为多个不同的 VLAN，汇聚层设备需要实现这些 VLAN 间的通信问题，同时，汇聚层设备还需对每一个接入层设备连接汇聚层设备的链路（上联链路）的流量实施控制。根据连接在不同 VLAN 的用户终端类型，汇聚层设备需要根据制定的安全策略，对 VLAN 之间的信息交换过程实施控制。

核心层设备实施路由功能时，必须实现 IP 分组的快速转发，实施交换功能时，必须实现 MAC 帧的快速转发，由于路由过程涉及 IP 分组分离、路由表查找、MAC 帧封装等操作，需要通过特定的机制实现 IP 分组的快速转发。

9.1.4 VLAN 互连实例

1. 单臂路由器实现 VLAN 互连实例

单臂路由器实现 VLAN 互连的网络结构如图 9.6 所示，终端 A、终端 B 和终端 G 分配给 VLAN 2，终端 E、终端 F 和终端 H 分配给 VLAN 3，终端 C 和终端 D 分配给 VLAN 4。由于每一个 VLAN 是一个独立的网络，属于不同 VLAN 的终端分配网络地址不同的 IP 地址，终端 IP 地址分配如图 9.6 所示。

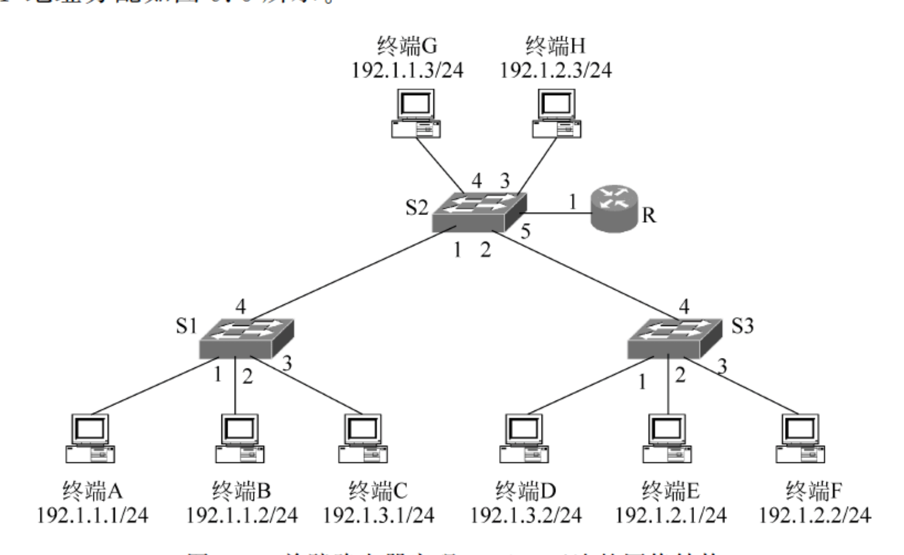


图 9.6 单臂路由器实现 VLAN 互连的网络结构

交换机 S2 端口 5 连接单臂路由器 R 的物理接口 1，连接单臂路由器物理接口的交换机端口必须满足以下条件：

- 是被 VLAN 2、VALN 3 和 VLAN 4 共享的共享端口，且是 802.1Q 标记端口；
- 交换式以太网中所有连接终端的端口与该端口之间存在交换路径。

之所以选择交换机 S2 端口 5 连接单臂路由器物理接口，是因为三个 VLAN 内的交换路径都经过交换机 S2，交换机 S2 为此已经创建了 VLAN 2、VLAN 3 和 VLAN 4，这样的话，一是很方便将交换机 S2 端口 5 配置成 VLAN 2、VALN 3 和 VLAN 4 的共享端口和 802.1Q 标记端口；二是很方便地建立交换机 S2 端口 5 与所有连接终端的交换机端口之间的交换路径。通过表 9.1 所示的 VLAN 端口配置，使得所有属于同一 VLAN 的终端之间存在交换路径，所有连接终端的交换机端口与交换机 S2 端口 5 之间存在交换路径。

表 9.1 VLAN 端口配置

交换机	VLAN 2		VLAN 3		VLAN 4	
	非标记端口	标记端口	非标记端口	标记端口	非标记端口	标记端口
交换机 S1	S1.1、S1.2	S1.4			S1.3	S1.4
交换机 S2	S2.4	S2.1、S2.5	S2.3	S2.2、S2.5		S2.1、S2.2、S2.5
交换机 S3			S3.2、S3.3	S3.4	S3.1	S3.4

路由器物理接口1被划分为三个逻辑接口，每一个逻辑接口绑定一个VLAN、分配IP地址和子网掩码，每一个逻辑接口分配的IP地址和子网掩码必须和与该接口绑定的VLAN的网络地址一致，同时，该逻辑接口的IP地址也成为连接在与该逻辑接口绑定的VLAN上的终端的默认网关地址。完成路由器物理接口划分和逻辑接口IP地址与子网掩码配置后，路由器自动建立如图9.7所示的路由表。

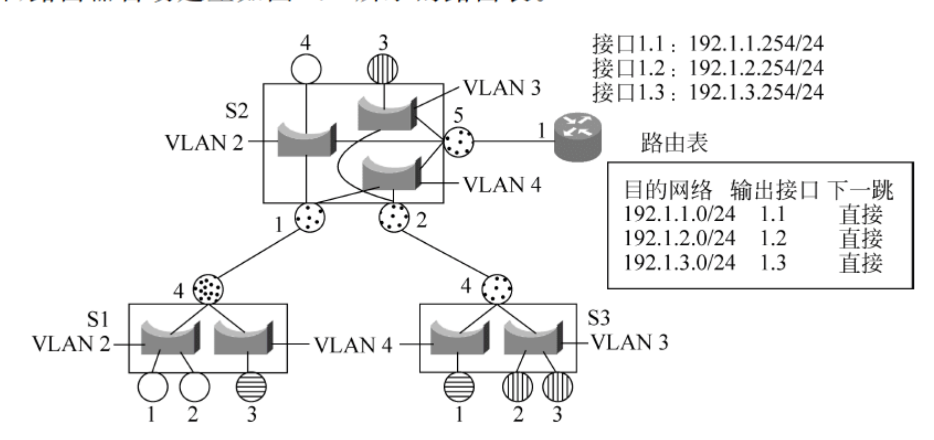


图9.7 单臂路由器对应的逻辑结构

完成上述配置后，图9.6中终端A与终端F的通信过程如下。

- 终端A根据配置的默认网关地址解析出路由器逻辑接口1.1的MAC地址。
- 终端A将源IP地址为192.1.1.1、目的IP地址为192.1.2.2的IP分组封装成以终端A的MAC地址为源MAC地址、以路由器逻辑接口1.1的MAC地址为目的MAC地址的MAC帧。
- 交换机S1根据该MAC帧的输入端口（端口1）确定该MAC帧所属的VLAN（VLAN 2），该MAC帧沿着交换机S1端口1至交换机S2端口5之间的属于VLAN 2的交换路径到达交换机S2端口5，由于交换机S2端口5是802.1Q标记端口，从该端口输出的MAC帧携带VLAN ID（VLAN 2）。
- 该MAC帧通过逻辑接口1.1进入路由器，路由器从中分离出IP分组，用IP分组的目的IP地址192.1.2.2检索路由器表，找到匹配的路由项，用目的IP地址192.1.2.2解析出终端F的MAC地址，根据输出接口1.2确定该逻辑接口绑定的VLAN（VLAN 3），重新将IP分组封装成以逻辑接口1.2的MAC地址为源MAC地址、以终端F的MAC地址为目的MAC地址、以VLAN 3为VLAN ID的MAC帧，将该MAC帧发送给交换机S2端口5。
- 该MAC帧沿着交换机S2端口5至交换机S3端口3之间属于VLAN 3的交换路径到达交换机S3端口3，由于交换机S3端口3是非标记端口（接入端口），从该端口输出的MAC帧删除VLAN ID。

2. 三层交换机实现VLAN互连实例

用三层交换机实现VLAN互连的网络结构如图9.8所示，图中S2是三层交换机，S1和S3是二层交换机。终端A、终端B和终端G分配给VLAN 2，终端E、终端F和终端H分配给VLAN 3，终端C和终端D分配给VLAN 4。由于每一个VLAN是一个独立的网

络，属于不同 VLAN 的终端分配网络地址不同的 IP 地址，终端 IP 地址分配如图 9.8 所示。

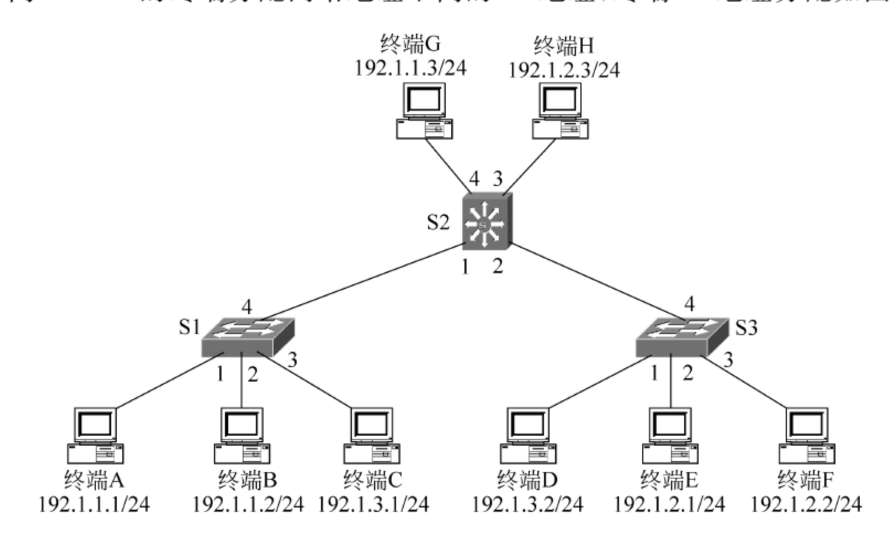


图 9.8 三层交换机实现 VLAN 互连的网络结构

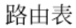
为了用三层交换机实现 VLAN 互连，必须在三层交换机 S2 中创建 VLAN 2、VLAN 3 和 VLAN 4，对每一个 VLAN，必须存在分配给该 VLAN 的交换机端口，交换机端口可以作为接入端口或共享端口分配给该 VLAN。其他交换机属于某个 VLAN 的端口与三层交换机中属于同一 VLAN 的端口之间必须建立交换路径。如交换机 S1 中属于 VLAN 2 的端口 1，与三层交换机 S2 中同样属于 VLAN 2 的端口 1 之间必须建立属于 VLAN 2 的交换路径。通过表 9.2 所示的 VLAN 端口配置，所有属于同一 VLAN 的终端之间存在交换路径，其他交换机属于某个 VLAN 的端口与三层交换机中属于同一 VLAN 的端口之间存在交换路径。

表 9.2 VLAN 端口配置

交换机	VLAN 2		VLAN 3		VLAN 4	
	非标记端口	标记端口	非标记端口	标记端口	非标记端口	标记端口
交换机 S1	S1.1、S1.2	S1.4			S1.3	S1.4
三层交换机 S2	S2.4	S2.1	S2.3	S2.2		S2.1、S2.2
交换机 S3			S3.2、S3.3	S3.4	S3.1	S3.4

注：S1.1 表示交换机 S1 的端口 1。

为每一个 VLAN 定义 IP 接口，为 IP 接口分配 IP 地址和子网掩码，每一个 IP 接口分配的 IP 地址和子网掩码必须与该 IP 接口对应的 VLAN 的网络地址一致。同时，该 IP 接口的 IP 地址也成为连接在与该 IP 接口对应的 VLAN 上终端的默认网关地址。完成 IP 接口定义和 IP 接口 IP 地址与子网掩码配置后，三层交换机 S2 自动建立图 9.9 所示的路由表。值得强调的是，三层交换机的路由模块通过三层交换机内部背板实现与三层交换机其他功能模块之间的通信，因此二层交换路径与 IP 接口及路由模块之间的传输通道对用户是透明的。这和单臂路由器互连 VLAN 方式需要用外部物理链路互连单臂路由器物理接口与交换机中被所有 VLAN 共享的共享端口是不同的。



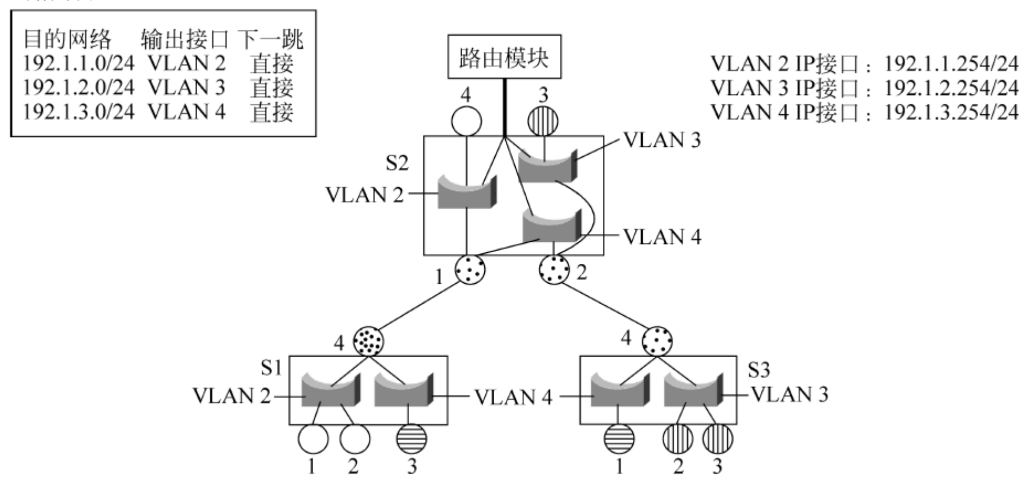


图 9.9　三层交换机对应的逻辑结构

9.2　三层交换过程

9.2.1　三层交换机结构

图 9.10 所示三层交换机由路由模块、交换结构、路由表、二层转发表和三层转发表组成，路由模块的功能有三，一是运行路由协议，通过和其他三层交换机交换路由消息构建路由表。二是用 IP 分组的目的 IP 地址检索路由表，找到匹配路由项，根据该路由项确定输出端口和下一跳结点的 IP 地址，通过 ARP 地址解析过程获取下一跳结点的 MAC 地址。三是根据 IP 分组的源 IP 地址、目的 IP 地址、封装 IP 分组的 MAC 帧的源 MAC 地址、MAC 帧输出端口、下一跳结点的 MAC 地址等信息构建三层转发表。

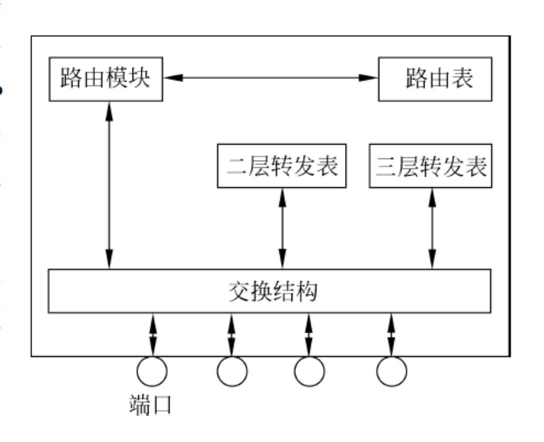


图 9.10　三层交换机结构

路由表中的每一项路由项用于指明通往某个网络的传输路径，它的格式是＜目的网络，距离，VLAN，下一跳结点地址＞。之所以用 VLAN 代替输出端口是因为三层交换机中所有属于某个 VLAN 的端口都有可能成为输出端口。因此，通过路由项只能确定输出端口所属的 VLAN。在通过 ARP 地址解析过程确定下一跳结点的 MAC 地址后，通过在二层转发表检索与该 MAC 地址匹配的转发项，才能确定输出端口。

二层转发表中的每一项转发项的功能有两个方面：一是用于指定端口所属的 VLAN；二是用于指明通往目的终端的传输路径。它的格式是＜VLAN，MAC 地址，端口＞，VLAN 字段给出端口所属的 VLAN，MAC 地址字段给出目的终端的 MAC 地址，端口字段给出以

MAC 地址字段值为目的 MAC 地址的 MAC 帧的输出端口。

路由模块完成 IP 分组路由操作的过程如下：如果需要路由模块完成 IP 分组的路由功能，封装该 IP 分组的 MAC 帧的目的 MAC 地址是表明接收端是三层交换机路由模块的特殊 MAC 地址，交换结构将以这样的 MAC 地址为目的 MAC 地址的 MAC 帧转发给路由模块。路由模块从中分离出 IP 分组，用该 IP 分组的目的 IP 地址检索路由表，根据匹配的路由项确定下一跳结点的 IP 地址和输出端口所属的 VLAN，通过 ARP 地址解析过程获取下一跳结点的 MAC 地址，将该 MAC 帧重新封装成以表明发送端是路由模块的特殊 MAC 地址为源 MAC 地址、以下一跳结点 MAC 地址为目的 MAC 地址、以输出端口所属 VLAN 为 VLAN ID 的 MAC 帧，通过检索二层转发表确定输出端口，通过输出端口输出该 MAC 帧。

三层转发表是三层交换机特有的，它的作用是以二层交换的方式实现三层路由功能。三层转发表对应特定的目的 IP 地址，记录下一跳结点的 MAC 地址、表示三层交换机路由模块的特定 MAC 地址、记录输出端口及输出端口所属的 VLAN，如果以后继续接收到相同目的 IP 地址的 IP 分组，无需路由模块进行路由操作，直接通过目的 IP 地址检索三层转发表，重新获取将该 IP 分组封装成 MAC 帧所需的全部信息和用于输出该重新封装的 MAC 帧的端口。三层转发表中每一项三层转发项的格式是＜目的 IP 地址，源 MAC 地址，目的 MAC 地址，VLAN，输出端口＞，源 MAC 地址是用于表明路由模块的特殊 MAC 地址，目的 MAC 地址是下一跳结点的 MAC 地址，VLAN 是输出端口所属的 VLAN。当交换结构接收到以表明接收端是该三层交换机路由模块的特殊 MAC 地址为目的 MAC 地址的 MAC 帧时，分离出 IP 分组，以该 IP 分组的目的 IP 地址检索三层转发表，如果找到匹配的三层转发项，重新将 IP 分组封装成以该三层转发项中的源和目的 MAC 地址为源和目的 MAC 地址、以该三层转发项中的 VLAN 为 VLAN ID 的 MAC 帧，通过该三层转发项中的输出端口输出重新封装的 MAC 帧。只有当三层转发表中检索不到与该目的 IP 地址匹配的三层转发项时，才将该 MAC 帧转发给路由模块。

交换结构首先根据 MAC 帧的目的 MAC 地址确定是二层交换操作，还是三层路由操作，对于实施二层交换操作的 MAC 帧，首先完成地址学习过程，然后通过检索二层转发表确定输出端口，通过输出端口输出该 MAC 帧。对于实施三层路由操作的 MAC 帧，分离出 IP 分组，通过三层地址学习过程创建三层转发项，三层转发项中的目的 IP 地址字段＝该 IP 分组的源 IP 地址，源 MAC 地址＝该 MAC 帧的目的 MAC 地址，目的 MAC 地址＝该 MAC 帧的源 MAC 地址，VLAN＝该 MAC 帧的 VLAN ID，输出端口＝该 MAC 帧的接收端口。交换结构通过接收到的以表明接收端是该三层交换机路由模块的特殊 MAC 地址为目的 MAC 地址的 MAC 帧构建对应三层转发项的过程称为三层地址学习过程。交换结构完成三层地址学习过程后，用 IP 分组的目的 IP 地址检索三层转发项，如果找到匹配的三层转发项，直接通过交换结构完成 IP 分组转发过程，如果找不到匹配的三层转发项，将该 MAC 帧转发给路由模块。

9.2.2 二层交换过程

图 9.11(a)给出终端与三层交换机之间的连接过程，三层交换机 4 个端口分别连接 4 个终端，其中端口 1 和端口 3 分配给 VLAN 2，端口 2 和端口 4 分配给 VLAN 3，端口与 VLAN 之间的关系如表 9.3 所示，图 9.11(b)是图 9.11(a)对应的逻辑结构。连接在属于不同 VLAN 的交换机端口的终端，必须分配网络地址不同的 IP 地址，如为连接在交换机端

口1的终端A分配IP地址和子网掩码192.1.1.1/24,并因此计算出终端A所在网络的网络地址是192.1.1.0/24。为连接在交换机端口2的终端B分配IP地址和子网掩码192.1.2.1/24,并因此计算出终端B所在网络的网络地址是192.1.2.0/24。连接在属于相同VLAN的交换机端口的终端,必须分配网络地址相同的IP地址。如分别为终端A和终端C分配IP地址和子网掩码192.1.1.1/24和192.1.1.2/24,因此计算出它们的网络地址都是192.1.1.0/24。

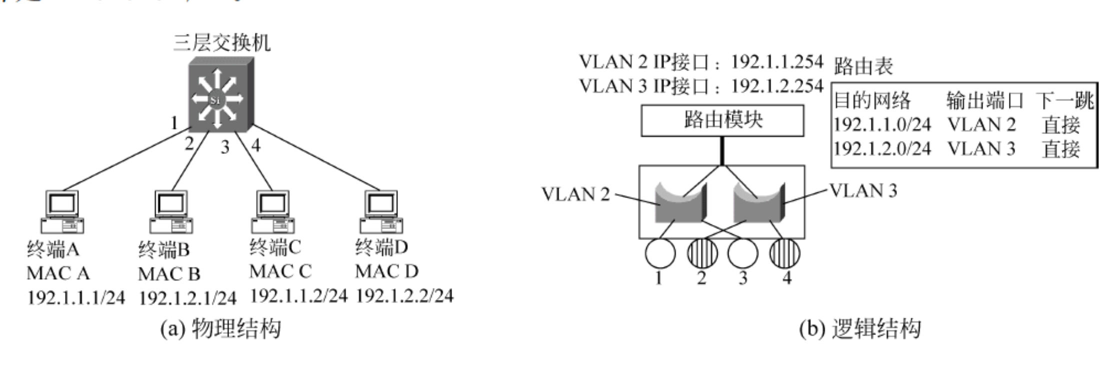


图9.11 网络结构

二层交换只能实现两个连接在属于相同VLAN的交换机端口的终端之间的MAC帧传输过程。终端A向终端C发送MAC帧前,必须获取终端C的MAC地址,然后构建以终端A的MAC地址(MAC A)为源MAC地址、终端C的MAC地址(MAC C)为目的MAC地址的MAC帧,并向三层交换机发送该MAC帧。

三层交换机通过端口1接收到该MAC帧后,确定该MAC帧的目的MAC地址不是表示接收端为该三层交换机路由模块的特殊MAC地址,完成该MAC帧的地址学习过程,在二层转发表中创建VLAN=VALN 2,MAC地址=MAC A,端口=端口1的转发项。然后在二层转发表中检索MAC地址与MAC C匹配的转发项,在没有找到与MAC C匹配的转发项的情况下,通过其他所有属于VLAN 2的交换机端口转发该MAC帧。这里由于属于VLAN 2的端口只有端口1和端口3,除了输入端口——端口1外,属于VLAN 2的其他端口只有端口3,交换结构通过端口3输出该MAC帧。

当终端C向终端A回送MAC帧时,终端C构建以终端C的MAC地址(MAC C)为源MAC地址、终端A的MAC地址(MAC A)为目的MAC地址的MAC帧,并向三层交换机发送该MAC帧。三层交换机通过端口3接收到该MAC帧后,确定该MAC帧的目的MAC地址不是表示接收端为该三层交换机路由模块的特殊MAC地址,完成该MAC帧的地址学习过程,在二层转发表中创建VLAN=VALN 2,MAC地址=MAC C,端口=端口3的转发项。完成上述转发项创建后的二层转发表如表9.4所示。然后在二层转发表中检索MAC地址与MAC A匹配的转发项,如果找到MAC地址与MAC A相同的转发项,通过该转发项指定的输出端口——端口1输出该MAC帧。

表9.3 三层交换机分配给不同VLAN的端口

VLAN	端口
VLAN 2	1,3
VLAN 3	2,4

表 9.4 二层转发表

VLAN	MAC 地址	端口
2	MAC A	1
2	MAC C	3

9.2.3 三层路由过程

1. 三层交换机初始配置

三层交换机进行三层路由操作前，必须定义 IP 接口，为 IP 接口分配 IP 地址和子网掩码，这里需要三层交换机分别对应 VLAN 2 和 VLAN 3 定义两个 IP 接口，分别为这两个 IP 接口分配 IP 地址和子网掩码 192.1.1.254/24 和 192.1.2.254/24。根据分配给这两个 IP 接口的 IP 地址和子网掩码计算出的网络地址必须与该 IP 接口对应的 VLAN 的网络地址一致。这两个 IP 接口的 IP 地址也成为分别连接在属于 VLAN 2 和 VLAN 3 的交换机端口上的终端的默认网关地址。这意味着终端 A 和终端 C 的默认网关地址为 192.1.1.254，终端 B 和终端 D 的默认网关地址为 192.1.2.254。完成 IP 接口定义和 IP 地址分配后，三层交换机创建如表 9.5 所示的路由表。

表 9.5 三层交换机路由表

目的网络	输出端口	下一跳	距离
192.1.1.0/24	VLAN 2	直接	0
192.1.2.0/24	VLAN 3	直接	0

三层交换机需要指定用于表明接收端是该三层交换机的路由模块的特殊 MAC 地址，不同厂家有着不同的指定该特殊 MAC 地址的方式，这里假定每一个三层交换机端口有着唯一的 MAC 地址，所以以某个三层交换机端口的 MAC 地址为目的 MAC 地址的 MAC 帧都是需要转发给路由模块的 MAC 帧。图 9.11 中四个三层交换机端口(端口 1～端口 4)对应的 MAC 地址分别是 MAC R1～MAC R4。

2. 路由 IP 分组过程

当终端 A 向终端 B 发送 IP 分组时，必须先获取终端 B 的 IP 地址 192.1.2.1，根据终端 A 配置的子网掩码 255.255.255.0，求出终端 A 和终端 B 所在网络的网络地址分别是 192.1.1.0/24 和 192.1.2.0/24。由于终端 A 和终端 B 连接在不同的网络，终端 A 需要先将 IP 分组发送给默认网关。为了获取默认网关的 MAC 地址，终端 A 广播 ARP 请求报文，请求解析 IP 地址 192.1.1.254 对应的 MAC 地址。该 ARP 请求报文在 VLAN 2 中广播，到达所有连接在属于 VLAN 2 的交换机端口的终端和路由模块。路由模块将接收该 ARP 请求报文的端口的 MAC 地址(MAC R1)作为 VLAN 2 对应的 IP 接口的 MAC 地址，将包含 IP 地址(192.1.1.254)和 MAC 地址(MAC R1)的 ARP 响应报文发送给终端 A。终端 A 构建源 IP 地址为 192.1.1.1、目的 IP 地址为 192.1.2.1 的 IP 分组，将该 IP 分组封装成以 MAC A 为源 MAC 地址、以 MAC R1 为目的 MAC 地址的 MAC 帧，将该 MAC 帧发送给三

层交换机。三层交换机交换结构通过端口 1 接收到该 MAC 帧，根据目的 MAC 地址(MAC R1)确定该 MAC 帧的接收端是路由模块，从 MAC 帧中分离出 IP 分组。首先通过三层地址学习过程创建一项三层转发项，其中目的 IP 地址为 IP 分组源 IP 地址(192.1.1.1)、源 MAC 地址为 MAC 帧目的 MAC 地址(MAC R1)、目的 MAC 地址为 MAC 帧源 MAC 地址(MAC A)、输出端口为接收该 MAC 帧端口——端口 1。由于端口 1 是接入端口，从该端口输出的 MAC 帧无需携带 VLAN ID，因此，三层转发项中的 VLAN ID 为空白。根据三层地址学习过程创建的三层转发项内容如表 9.6 所示。完成三层地址学习过程后，用 IP 分组的目的地址 192.1.2.1 检索三层转发表，没有找到与 IP 地址 192.1.2.1 匹配的三层转发项，交换结构将该 MAC 帧转发给路由模块。

路由模块分离出 IP 分组，用目的 IP 地址 192.1.2.1 检索路由表，找到匹配的路由项，确定输出端口所属的 VLAN 是 VALN 3、下一跳结点是目的终端自身。通过在 VLAN 3 广播 ARP 请求报文，获取终端 B 的 MAC 地址(MAC B)，用 MAC B 检索二层转发表确定输出端口是端口 2 且端口 2 是接入端口。路由模块一方面将 IP 分组封装成以交换机端口 2 的 MAC 地址(MAC R2)为源 MAC 地址、终端 B 的 MAC 地址为目的 MAC 地址的 MAC 帧，并将 MAC 帧通过端口 2 输出。另一方面，创建一项三层转发项，其中目的 IP 地址为 IP 分组目的 IP 地址 192.1.2.1、源 MAC 地址为交换机端口 2 的 MAC 地址(MAC R2)、目的 MAC 地址为终端 B 的 MAC 地址(MAC B)、输出端口为端口 2。由于端口 2 是接入端口，从该端口输出的 MAC 帧无须携带 VLAN ID，因此，三层转发项中的 VLAN ID 为空白。该三层转发项内容如表 9.6 所示。从端口 2 输出的 MAC 帧到达终端 B。

表 9.6　三层转发表

目的 IP 地址	源 MAC 地址	目的 MAC 地址	输出端口	VLAN ID
192.1.1.1	MAC R1	MAC A	1	—
192.1.2.1	MAC R2	MAC B	2	—

如果终端 B 向终端 A 发送 IP 分组，终端 B 构建源 IP 地址为 192.1.2.1、目的 IP 地址为 192.1.1.1 的 IP 分组，将该 IP 分组封装成以终端 B 的 MAC 地址为源 MAC 地址、交换机端口 2 的 MAC 地址(MAC R2)为目的 MAC 地址的 MAC 帧，将该 MAC 帧发送给三层交换机，三层交换机交换结构接收到该 MAC 帧后，根据目的 MAC 地址 MAC R2 确定该 MAC 帧的接收端是路由模块，从该 MAC 帧分离出 IP 分组，用 IP 分组的目的 IP 地址检索三层转发表，找到匹配的三层转发项，直接根据该三层转发项的信息重新将该 IP 分组封装成 MAC 帧，重新封装的 MAC 帧的源 MAC 地址是三层转发项中的源 MAC 地址(MAC R1)、目的 MAC 地址是三层转发项的目的 MAC 地址(MAC A)。将重新封装的 MAC 帧通过三层转发项指定的输出端口——端口 1 输出。从端口 1 输出的 MAC 帧到达终端 A。

后续终端 A 发送给终端 B 的 IP 分组，由于可以在三层转发表中检索到与目的 IP 地址 192.1.2.1 匹配的三层转发项，因此直接根据该三层转发项重新封装该 IP 分组，并将重新封装的 MAC 帧通过三层转发项指定的输出端口输出，无须经过路由模块的路由操作。

和二层转发项一样，每一项三层转发项关联一个定时器，每当经过该三层转发项完成 IP 分组转发操作时，刷新该定时器，一旦定时器溢出，从三层转发表中删除该三层转发项。

9.3 三层交换机应用方式

9.3.1 IP接口集中到单个三层交换机

1. IP接口配置和网络逻辑结构

互连网络结构如图9.12所示，S1和S2是三层交换机，要求终端A和终端C属于VLAN 2，终端B和终端D属于VLAN 3，可以通过在单个三层交换机上定义VLAN 2和VLAN 3对应的IP接口，实现属于同一VLAN的终端之间通信和属于不同VLAN的终端之间通信的功能。

将IP接口集中到单个三层交换机的逻辑结构如图9.13所示，由于VLAN 2和VLAN 3对应的IP接口定义在三层交换机S1中，因此，属于同一VLAN的终端之间必须建立交换路径，属于VLAN 2和VLAN 3的终端必须建立与三层交换机S1之间的交换路径。图9.14给出了交换机S1和交换机S2的VLAN配置，三层交换机S1作为三层交换机使用，定义分别对应VLAN 2和VLAN 3的IP接口，并为这两个IP接口分配IP地址和子网掩码，为这两个IP接口分配IP地址和子网掩码后，建立图9.14所示的路由表。三层交换机S2作为普通二层交换机使用，用于建立属于同一VLAN的终端之间的交换路径和连接在三层交换机S2上的终端与三层交换机S1之间的交换路径。交换机S1和交换机S2的VLAN端口配置如表9.7所示。

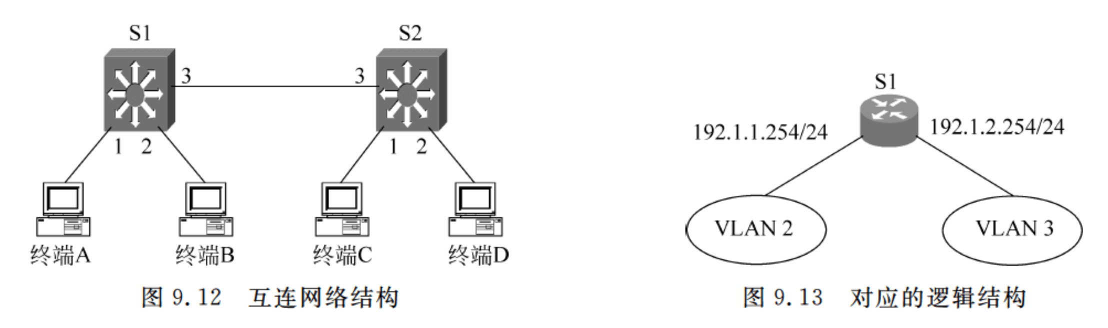


图9.12 互连网络结构

图9.13 对应的逻辑结构

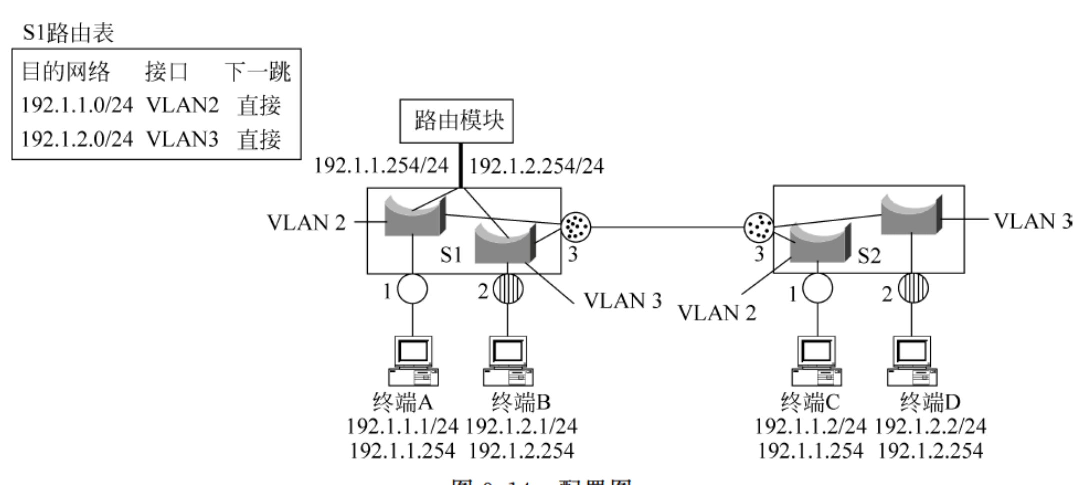


图9.14 配置图

表 9.7 VLAN 端口配置

交换机	VLAN 2		VLAN 3	
	非标记端口	标记端口	非标记端口	标记端口
交换机 S1	1.1	1.3	1.2	1.3
交换机 S2	2.1	2.3	2.2	2.3

注：1.1 表示交换机 S1 的端口 1。

2. VLAN 之间 IP 分组传输过程

假定三层交换机 S1 三个端口(端口 1～端口 3)的 MAC 地址分别是 MAC R1～MAC R3，终端 C 和终端 D 的 MAC 地址分别是 MAC C 和 MAC D。终端 C 至终端 D 的 IP 分组传输过程如下。

- 终端 C 通过地址解析过程获取路由模块的 MAC 地址(MAC R3)。
- 终端 C 构建以 IP 地址 192.1.1.2 为源 IP 地址、以 IP 地址 192.1.2.2 为目的 IP 地址的 IP 分组。
- 终端 C 将 IP 分组封装成以 MAC C 为源 MAC 地址、以 MAC R3 为目的 MAC 地址的 MAC 帧。
- 交换机 S2 通过二层交换过程从端口 3 输出该 MAC 帧，由于端口 3 是 802.1Q 标记端口，从端口 3 输出的 MAC 帧携带 VLAN ID——VLAN 2。
- 三层交换机 S1 通过端口 3 接收该 MAC 帧，根据该 MAC 帧的目的 MAC 地址确定该 MAC 帧的接收端是路由模块，从 MAC 帧中分离出 IP 分组，通过三层地址学习过程创建表 9.8 所示的 IP 地址 192.1.1.2 对应的三层转发项。
- 用 IP 分组的目的 IP 地址检索三层转发表，没有找到匹配的三层转发项，将该 MAC 帧转发给路由模块。
- 路由模块确定该 IP 分组的下一跳是目的终端自身、输出端口所属的 VLAN 是 VLAN 3，通过地址解析过程获取终端 D 的 MAC 地址(MAC D)，通过检索二层转发表确定输出端口——端口 3，将 IP 分组封装成以 MAC R3 为源 MAC 地址、MAC D 为目的 MAC 地址、VLAN ID 为 VLAN 3 的 MAC 帧，通过三层交换机 S1 端口 3 输出该 MAC 帧，同时在三层转发表中创建表 9.8 所示的 IP 地址 192.1.2.2 对应的三层转发项。
- 交换机 S2 通过端口 3 接收到该 MAC 帧，根据该 MAC 帧携带的 VLAN ID 确定转发该 MAC 帧的网桥，由 VLAN 3 对应的网桥通过二层交换过程将该 MAC 帧从端口 2 转发出去，完成 IP 分组终端 C 至终端 D 的传输过程。

表 9.8 三层交换机 S1 三层转发表

目的 IP 地址	源 MAC 地址	目的 MAC 地址	输出端口	VLAN ID
192.1.1.2	MAC R3	MAC C	3	VLAN 2
192.1.2.2	MAC R3	MAC D	3	VLAN 3

9.3.2 两个三层交换机同时定义所有 VLAN 对应的 IP 接口

1. IP 接口配置和网络逻辑结构

网络结构和终端与 VLAN 之间关系与 9.3.1 节相同，可以通过在三层交换机 S1 和三层交换机 S2 上同时定义 VLAN 2 和 VLAN 3 对应的 IP 接口，实现属于同一 VLAN 的终端之间通信和属于不同 VLAN 的终端之间通信的功能。

在两个三层交换机上同时定义 VLAN 2 和 VLAN 3 对应的 IP 接口的逻辑结构如图 9.15 所示，由于三层交换机 S1 和三层交换机 S2 中同时定义 VLAN 2 和 VLAN 3 对应的 IP 接口，因此，属于同一 VLAN 的终端之间必须建立交换路径，属于 VLAN 2 和 VLAN 3的终端必须建立与三层交换机 S1 和三层交换机 S2 之间的交换路径。

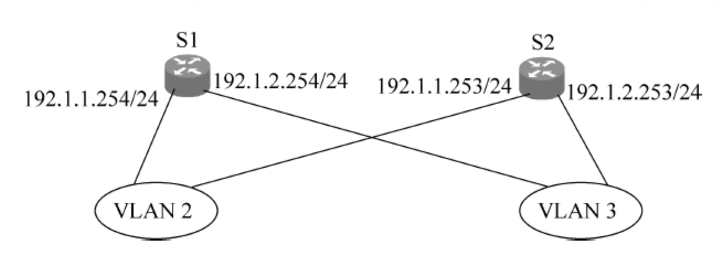


图 9.15 逻辑结构

图 9.16 给出了三层交换机 S1 和三层交换机 S2 的 VLAN 配置，三层交换机 S1 和三层交换机 S2 均作为三层交换机使用，分别定义对应 VLAN 2 和 VLAN 3 的 IP 接口，并为 IP 接口分配 IP 地址和子网掩码。完成 IP 接口定义与 IP 地址和子网掩码分配后，建立如图 9.16 所示的路由表。三层交换机 S1 和三层交换机 S2 同时具有普通二层交换机功能，用于建立属于同一 VLAN 的终端之间交换路径和连接在一个三层交换机上的终端与另一个三层交换机之间的交换路径。交换机 S1 和交换机 S2 的 VLAN 配置如表 9.7 所示。对于图 9.16 所示的 IP 接口配置，属于 VLAN 2 的终端可以任意选择 192.1.1.254 或 192.1.1.253作为默认网关地址，同样，属于 VLAN 3 的终端可以任意选择 192.1.2.254 或 192.1.2.253 作为默认网关地址。

2. VLAN 之间 IP 分组传输过程

假定三层交换机 S1 三个端口的 MAC 地址分别为 MAC R11～MAC R13，三层交换机 S2 三个端口的 MAC 地址分别为 MAC R21～MAC R23，终端 C 的 MAC 地址为 MAC C，终端 D 的 MAC 地址为 MAC D。根据图 9.16 所示的 IP 接口配置和终端 C 与终端 D 选择的默认网关地址，终端 C 至终端 D 的 IP 分组传输过程如下。

- 由于终端 C 选择的默认网关地址是三层交换机 S1 中对应 VLAN 2 的 IP 接口的 IP 地址，因此，终端 C 通过对默认网关地址 192.1.1.254 进行地址解析，得到对应的 MAC 地址 MAC R13。
- 终端 C 发送给默认网关的 MAC 帧沿着三层交换机 S2 连接终端 C 的端口至三层交换机 S1 端口 3 的交换路径进入三层交换机 S1。
- 三层交换机 S1 的交换结构通过三层地址学习过程创建如表 9.9 所示的与 IP 地址

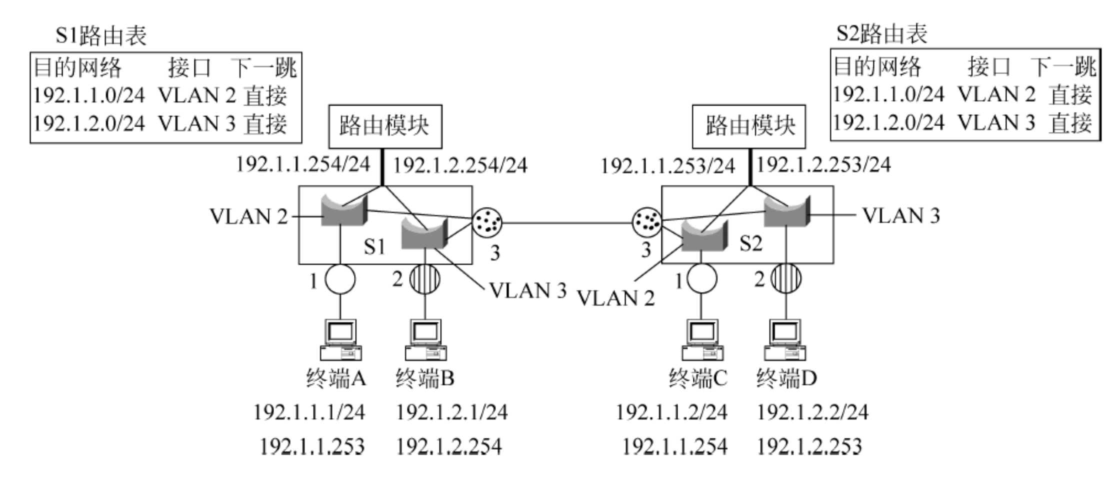


图 9.16　配置图

192.1.1.2 对应的三层转发项，同时，将该 MAC 帧转发给路由模块。

- 三层交换机 S1 的路由模块完成路由表检索、下一跳结点 IP 地址解析等，创建如表 9.9 所示的与 IP 地址 192.1.2.2 对应的三层转发项，将重新封装后的 MAC 帧通过端口 3 输出。
- 重新封装后的 MAC 帧沿着三层交换机 S2 端口 3 至三层交换机 S1 连接终端 D 的端口的交换路径到达终端 D，完成 IP 分组终端 C 至终端 D 传输过程。

表 9.9　三层交换机 S1 三层转发表

目的 IP 地址	源 MAC 地址	目的 MAC 地址	输出端口	VLAN ID
192.1.1.2	MAC R13	MAC C	3	VLAN 2
192.1.2.2	MAC R13	MAC D	3	VLAN 3

终端 D 至终端 C 的 IP 分组传输过程如下。

- 由于终端 D 选择的默认网关地址是三层交换机 S2 中对应 VLAN 3 的 IP 接口的 IP 地址，因此，终端 D 通过对默认网关地址 192.1.2.253 进行地址解析，得到对应的 MAC 地址 MAC R22。
- 三层交换机 S2 的交换结构通过端口 2 接收到终端 D 发送给默认网关的 MAC 帧，确定该 MAC 帧的接收端是三层交换机 S2 的路由模块，通过三层地址学习过程创建如表 9.10 所示的与 IP 地址 192.1.2.2 对应的三层转发项，同时，将该 MAC 帧转发给路由模块。
- 三层交换机 S2 的路由模块完成路由表检索、下一跳结点 IP 地址解析等，创建如表 9.10 所示的与 IP 地址 192.1.1.2 对应的三层转发项，将重新封装后的 MAC 帧直接通过端口 1 传输给终端 C，完成 IP 分组终端 D 至终端 C 的传输过程。

表 9.10　三层交换机 S2 三层转发表

目的 IP 地址	源 MAC 地址	目的 MAC 地址	输出端口	VLAN ID
192.1.2.2	MAC R22	MAC D	2	—
192.1.1.2	MAC R21	MAC C	1	—

9.3.3 两个三层交换机分别定义两个 VLAN 对应的 IP 接口

1. IP 接口配置和网络逻辑结构

网络结构和终端与 VLAN 之间关系与 9.3.1 节相同，通过分别在三层交换机 S1 上定义 VLAN 2 对应的 IP 接口，在三层交换机 S2 上定义 VLAN 3 对应的 IP 接口，实现属于同一 VLAN 的终端之间通信，属于不同 VLAN 的终端之间通信的功能。

两个三层交换机分别定义两个 VLAN 对应的 IP 接口的逻辑结构如图 9.17 所示，其中 VLAN 2 直接和交换机 S1 相连，VLAN 3 直接和交换机 S2 相连，为了实现 VLAN 2 和 VLAN 3 之间通信，需要用 VLAN 4 互连交换机 S1 和交换机 S2。和两个路由器互连三个 VLAN 不同，VLAN 2 包含物理上连接在交换机 S2 上的终端 C，因此对于 VLAN 2 和终端 C，交换机 S2 是一个二层交换机，用于创建终端 C 至交换机 S1 中 VLAN 2 对应的 IP 接口和终端 A 之间的交换路径。同理，对于 VLAN 3 和终端 B，交换机 S1 是一个二层交换机，用于创建终端 B 至交换机 S2 中 VLAN 3 对应的 IP 接口和终端 D 之间的交换路径。这是三层交换机和路由器的本质区别，即三层交换机既可建立属于同一 VLAN 的终端之间的交换路径，又可建立不同 VLAN 之间的 IP 传输路径。

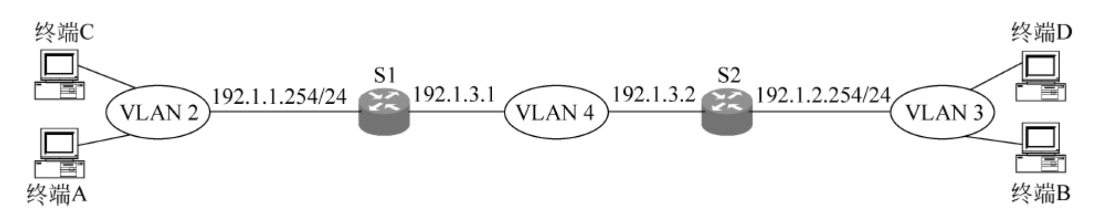


图 9.17 逻辑结构

交换机 S1 和交换机 S2 的 VLAN 配置如表 9.11 所示，这样配置是为了保证：①建立属于同一 VLAN 的终端之间的交换路径；②建立所有属于 VLAN 2 的终端至交换机 S1 的交换路径；③建立所有属于 VLAN 3 的终端至交换机 S2 的交换路径。同时，通过 VLAN 4 建立交换机 S1 中 VLAN 2 对应的 IP 接口至交换机 S2 中 VLAN 3 对应的 IP 接口之间的交换路径。为此，交换机 S1 需配置两个分别对应 VLAN 2 和 VLAN 4 的 IP 接口，为这两个 IP 接口分配 IP 地址和子网掩码，为 VLAN 2 对应的 IP 接口分配的 IP 地址和子网掩码既确定了 VLAN 2 的网络地址，同时又确定了连接在 VLAN 2 中的终端的默认网关地址，同样，交换机 S2 需配置两个分别对应 VLAN 3 和 VLAN 4 的 IP 接口，为这两个 IP 接口分配 IP 地址和子网掩码，为 VLAN 3 对应的 IP 接口分配的 IP 地址和子网掩码既确定了 VLAN 3 的网络地址，同时又确定了连接在 VLAN 3 中的终端的默认网关地址。交换机 S1 和交换机 S2 中为 VLAN 4 对应的 IP 接口分配的 IP 地址必须属于同一网络地址，对于交换机 S1，交换机 S2 中 VLAN 4 对应的 IP 接口的 IP 地址就是交换机 S1 通往 VLAN 3 的传输路径的下一跳地址，同样，对于交换机 S2，交换机 S1 中 VLAN 4 对应的 IP 接口的 IP 地址就是交换机 S2 通往 VLAN 2 的传输路径的下一跳地址。图 9.18 给出了 IP 接口配置及对应的交换机 S1 和交换机 S2 的路由表。

表 9.11　VLAN 端口配置

交换机	VLAN 2		VLAN 3		VLAN 4	
	非标记端口	标记端口	非标记端口	标记端口	非标记端口	标记端口
交换机 S1	1.1	1.3	1.2	1.3		1.3
交换机 S2	2.1	2.3	2.2	2.3		2.3

注：1.1 表示交换机 S1 的端口 1。

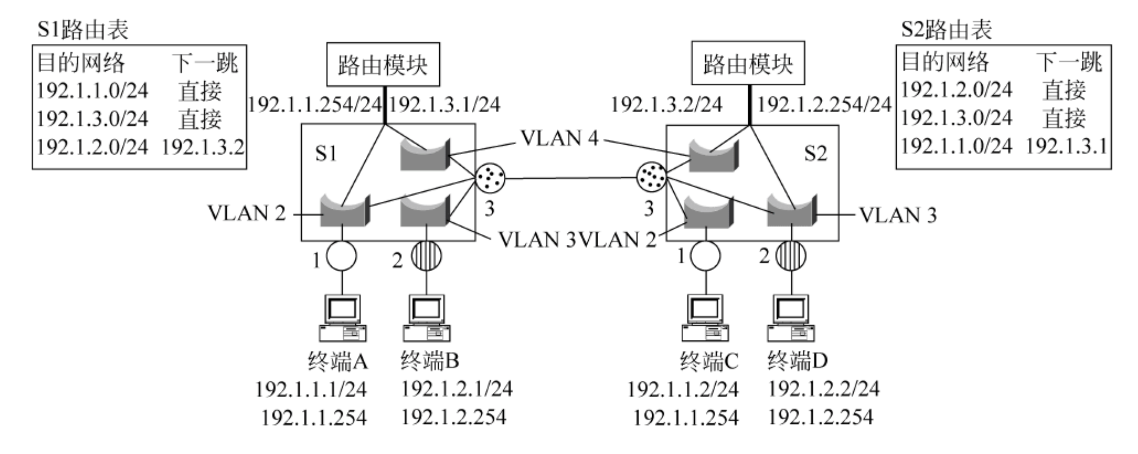


图 9.18　配置图

2. VLAN 之间 IP 分组传输过程

假定三层交换机 S1 三个端口的 MAC 地址分别为 MAC R11～MAC R13，三层交换机 S2 三个端口的 MAC 地址分别为 MAC R21～MAC R23，终端 C 的 MAC 地址为 MAC C，终端 D 的 MAC 地址为 MAC D。根据图 9.18 所示的 IP 接口配置和三层交换机 S1、三层交换机 S2 中的路由表内容，终端 C 至终端 D 的 IP 分组传输过程如下。

- 由于终端 C 选择的默认网关地址是三层交换机 S1 中对应 VLAN 2 的 IP 接口的 IP 地址，因此，终端 C 通过对默认网关地址 192.1.1.254 进行地址解析，得到对应的 MAC 地址 MAC R13。
- 终端 C 发送给默认网关的 MAC 帧沿着三层交换机 S2 连接终端 C 的端口至三层交换机 S1 端口 3 的交换路径进入三层交换机 S1。
- 三层交换机 S1 的交换结构根据目的 MAC 地址 MAC R13 确定该 MAC 帧的接收端是路由模块，通过三层地址学习过程创建如表 9.12 所示的与 IP 地址 192.1.1.2 对应的三层转发项，同时，将该 MAC 帧转发给路由模块。
- 三层交换机 S1 的路由模块从该 MAC 帧中分离出 IP 分组，用该 IP 分组的目的 IP 地址 192.1.2.2 检索路由表，找到匹配的路由项，确定输出端口所属的 VLAN 是 VLAN 4，下一跳结点地址是 192.1.3.2，对 IP 地址 192.1.3.2 进行地址解析，获得对应的 MAC 地址（MAC R23），用 MAC R23 检索二层转发表，确定输出端口是端口 3，路由模块根据以上信息创建如表 9.12 所示的与 IP 地址 192.1.2.2 对应的三层转发项，重新将 IP 分组封装成以 MAC R13 为源 MAC 地址、MAC R23 为目的

MAC 地址、VLAN ID 为 VLAN 4 的 MAC 帧，通过端口 3 输出该 MAC 帧。

- 该 MAC 帧沿着三层交换机 S1 端口 3 至三层交换机 S2 端口 3 的交换路径到达三层交换机 S2，三层交换机 S2 的交换结构根据目的 MAC 地址（MAC R23）确定该 MAC 帧的接收端是路由模块，通过三层地址学习过程创建如表 9.13 所示的与 IP 地址 192.1.1.2 对应的三层转发项，同时，将该 MAC 帧转发给路由模块。
- 三层交换机 S2 的路由模块从该 MAC 帧中分离出 IP 分组，用该 IP 分组的目的 IP 地址 192.1.2.2 检索路由表，找到匹配的路由项，确定输出端口所属的 VLAN 是 VLAN 3，下一跳结点是目的终端自身，对 IP 地址 192.1.2.2 进行地址解析，获得对应的 MAC 地址（MAC D），用 MAC D 检索二层转发表，确定输出端口是端口 2。路由模块根据以上信息创建如表 9.13 所示的与 IP 地址 192.1.2.2 对应的三层转发项，重新将 IP 分组封装成以 MAC R22 为源 MAC 地址、MAC D 为目的 MAC 地址的 MAC 帧，通过端口 2 输出该 MAC 帧，该 MAC 帧到达终端 D，完成 IP 分组终端 C 至终端 D 的传输过程。

表 9.12 三层交换机 S1 三层转发表

目的 IP 地址	源 MAC 地址	目的 MAC 地址	输出端口	VLAN ID
192.1.1.2	MAC R13	MAC C	3	VLAN 2
192.1.2.2	MAC R13	MAC 23	3	VLAN 4

表 9.13 三层交换机 S2 三层转发表

目的 IP 地址	源 MAC 地址	目的 MAC 地址	输出端口	VLAN ID
192.1.1.2	MAC R23	MAC 13	3	VLAN 4
192.1.2.2	MAC R22	MAC D	2	—

终端 D 至终端 C 的 IP 分组传输过程如下。

- 终端 D 将 IP 分组封装成以 MAC 地址（MAC D）为源 MAC 地址、MAC R22 为目的 MAC 地址的 MAC 帧，将该 MAC 帧通过端口 2 传输给三层交换机 S2。
- 三层交换机 S2 的交换结构根据目的 MAC 地址（MAC R22）确定该 MAC 帧的接收端是路由模块，从 MAC 帧中分离出 IP 分组，用 IP 分组的目的 IP 地址 192.1.1.2 检索三层转发表，找到匹配的三层转发项，根据表 9.13 中与目的 IP 地址 192.1.1.2 对应的三层转发项内容直接将 IP 分组封装成以 MAC R23 为源 MAC 地址、MAC R13 为目的 MAC 地址、VLAN ID 为 VLAN 4 的 MAC 帧，将该 MAC 帧通过端口 3 输出。
- 三层交换机 S1 通过端口 3 接收到该 MAC 帧，三层交换机 S1 的交换结构根据目的 MAC 地址（MAC R13）确定该 MAC 帧的接收端是路由模块，从 MAC 帧中分离出 IP 分组，用 IP 分组的目的 IP 地址 192.1.1.2 检索三层转发表，找到匹配的三层转发项，根据表 9.12 中与目的 IP 地址 192.1.1.2 对应的三层转发项内容直接将 IP 分组封装成以 MAC R13 为源 MAC 地址、MAC C 为目的 MAC 地址、VLAN ID 为 VLAN 2 的 MAC 帧，将该 MAC 帧通过端口 3 输出。
- 三层交换机 S2 通过端口 3 接收到该 MAC 帧，根据该 MAC 帧携带的 VLAN ID 将

该 MAC 帧提交给 VLAN 2 对应的网桥转发，VLAN 2 对应的网桥通过用 MAC C 检索二层转发表，确定该 MAC 帧的输出端口是端口 1，由于端口 1 是接入端口，删除该 MAC 帧的 VLAN ID，将该 MAC 帧通过端口 1 传输给终端 C，完成 IP 分组终端 D 至终端 C 的传输过程。

值得强调的是，不同厂家有着不同的三层转发表内容和三层转发表的创建方式，但基于三层交换机用于实现 VLAN 间 IP 分组转发这一特殊性，三层交换机一般都会通过创建三层转发表的方式，以交换手段实现 IP 分组不同 VLAN 间的转发过程，以此提高三层交换机的 IP 分组转发性能。

习题

9.1 VLAN 之间通信需要通过路由器或三层交换机是物理限制，还是逻辑限制？它和分别连接在以太网和 ATM 网络上的两个终端之间通信必须经过路由器的原因有何异同？

9.2 简述路由器和三层交换机的区别。

9.3 简述三层交换机集二层交换和三层路由于一体的原因。

9.4 三层交换机对某个 MAC 帧进行二层交换操作或三层路由操作的依据是什么？

9.5 三层交换机的 IP 接口和路由器的逻辑接口有何异同？

9.6 简述通过三层地址学习过程创建三层转发项的过程。

9.7 三层转发表与路由表有何区别和关联？

9.8 给出图 9.19 中的 VLAN 划分和 IP 接口配置，并给出终端 A→终端 E 及终端 A→终端 D 之间的通信过程。要求：两个交换机都必须作为三层交换机使用。

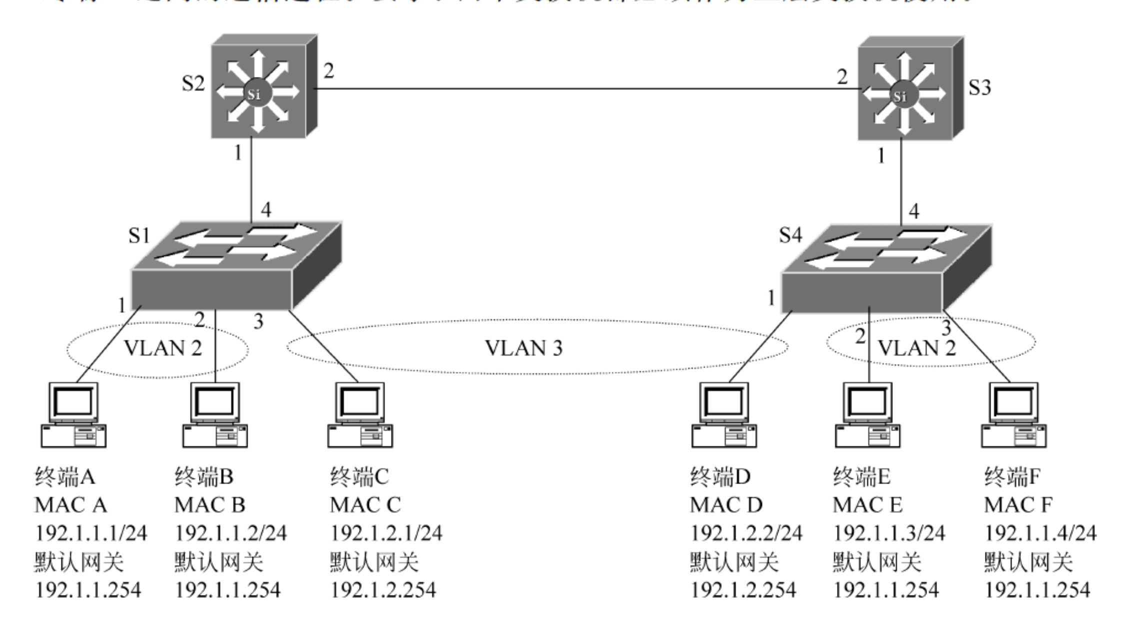


图 9.19 题 9.8 图

9.9 网络结构如图 9.20 所示,要求三个交换机(S1、S2 和 S3)都作为三层交换机使用,给出三层交换机的配置,包括 VLAN 划分、IP 接口定义、IP 接口 IP 地址和子网掩码分配、三层交换机路由表等。简述同一 VLAN 内终端之间和属于不同 VLAN 的终端之间的通信过程。

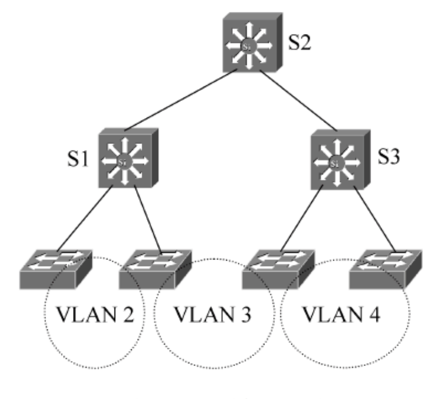


图 9.20 题 9.9 图

第10章 IPv6

以 IPv4 为基础的 Internet 在过去十多年间得到了飞速发展，Internet 的规模和应用方式发生了巨大的变化，面对庞大的规模和多种多样的应用方式，以 IPv4 为基础的 Internet 开始面临各种各样的问题，IPv4 的局限性开始显现。虽然人们提出了多种用于弥补 IPv4 缺陷的方法，但这些方法治标不治本。为了适应 Internet 的飞速发展，必须提出一种新的用于实现网络互连的协议，它就是 IPv6。

10.1 IPv4 的缺陷

10.1.1 地址短缺问题

IPv4 用 32 位二进制数表示 IP 地址，虽说 32 位二进制数能够提供四十多亿个 IP 地址，可满足全世界 2/3 人口的上网需求，但实际能够使用的地址空间远没有那么多，这是因为：

- IPv4 地址的分层结构导致大量地址空间被浪费，虽然无分类编址（CIDR）极大地缓解了这一问题，但浪费地址空间的现象依然存在。
- 保留的 E 类地址和用作组播的 D 类地址占用了近 12%的地址空间。
- 一些无法分配给网络终端的特殊地址，如主机号全 0 或全 1 的 IP 地址占用了近 2%的地址空间。

因此，随着 Internet 规模的不断扩大，地址短缺问题日益突出。目前普遍用于解决地址短缺问题的方法是网络地址转换（Network Address Translation，NAT）技术，正是无分类编址和 NAT 技术的出现，使得 IPv4 的地址短缺问题得到缓解，有一部分人甚至认为 IPv4 的地址短缺问题已经得到解决，这也是 IPv6 在很长一段时间内得不到重视，在未来很长一段时间内 Internet 仍然以 IPv4 为主的主要原因。但 NAT 也带来一些问题：一是破坏了 IP 的端到端通信模型，使得对等通信的双向会话变得困难。同时由于隐藏了源终端地址，使得一些需要在应用层 PDU 中给出源终端地址的应用难以实现。二是边界路由器需要记录大量的地址转换信息，这不仅对边界路由器的性能提出了更高要求，而且还影响网络性能。三是由于类似 IP Sec 这样的端到端安全功能不容许在传输过程中改变 IP 首部内容，因此，NAT 使类似 IP Sec 这样的端到端安全功能变得难以实现。四是 NAT 都是针对会话绑定地址映射，因此，一旦采用 NAT，必须先创建会话，这就将无连接的 IP 分组传输过程转变成了面向连接的传输过程。可以说无分类编址和 NAT 技术极大地缓解了 IPv4 的地址短缺

问题，使人们有更充分的时间来部署IPv6，但NAT只是权宜之计，不是解决IPv4地址短缺问题的根本方法。

随着无线局域网和个人数字助理(Personal Digital Assistant，PDA)的兴起，集移动通信和访问Internet资源的功能于一身的PDA将成为人们的首选，大量移动用户一旦成为Internet用户，地址短缺问题将立即成为亟待解决的紧迫问题。随着计算机和通信技术的发展，人们通过网络监测、控制家电已不是梦想，但这一切都是以家电成为网络终端设备为前提的，一旦大量家电需要接入Internet，地址短缺问题更是迫在眉睫。

10.1.2 复杂的分组首部

分组首部结构影响路由器转发分组的速率，目前通信链路的传输速率越来越高，10Gb/s的同步数字体系(Synchronous Digital Hierarchy，SDH)和以太网正成为主流广域网和局域网技术，在这种情况下，路由器实现线速转发越来越困难，路由器正日益成为网络性能的瓶颈。为了提高路由器转发速率，要求减少路由器转发分组所必须进行的操作，如差错检验。传统IPv4首部中有一首部检验和字段，路由器通过该字段检验IPv4分组首部在传输过程中是否发生错误，由于IPv4分组每经过一跳路由器，都会改变首部中TTL字段值，导致每一跳路由器都需要重新计算IPv4首部检验和字段值，增加了路由器转发IPv4分组所进行的操作。另一方面，随着通信技术的发展，通信链路的传输可靠性越来越高，传输出错的概率越来越小，而且链路层和传输层的差错控制足以检验出传输出错的分组，在网络层进行差错检验的必要性越来越小。在前面章节中也讲过，由于网络终端的处理能力越来越高，目前的趋势是尽量将处理功能转移到网络终端，以此简化路由器的转发处理，提高路由器转发分组的速率。因此，IPv4复杂的首部结构及与此对应的转发处理要求极大地限制了路由器转发IPv4分组的速率，也与目前尽量将处理功能转移到网络终端的趋势相悖。

10.1.3 QoS实现困难

设计IPv4时是无法想象到以它为基础的Internet能够支持目前的规模和应用方式，可以说IPv4的成功已经远远超出了设计者的预期。但随着统一网络的设想逐步得到实现，IPv4服务的缺陷也日益显现。虽然人们尽了很大的努力来弥补这一缺陷，但IPv4对分类服务的先天不足仍然严重制约了类似VOIP、IPTV等实时应用的开展。由于IPv4首部中没有用于标识流的流标签字段，路由器需要更多的处理能力对流进行分类，并在流分类的基础上提供分类服务。一方面加重了路由器的处理负担，影响路由器转发速率；另一方面也与目前尽量将处理功能转移到网络终端的趋势相悖。

10.1.4 安全机制先天不足

IPv4的设计目的是尽量方便进程间的通信，因此，并没有较多地考虑安全问题，但随着电子商务活动的日益频繁，信息资源的安全性越来越重要，需要总体上对网络安全进行设计，而不是软件补丁似的头痛医头、脚痛医脚。

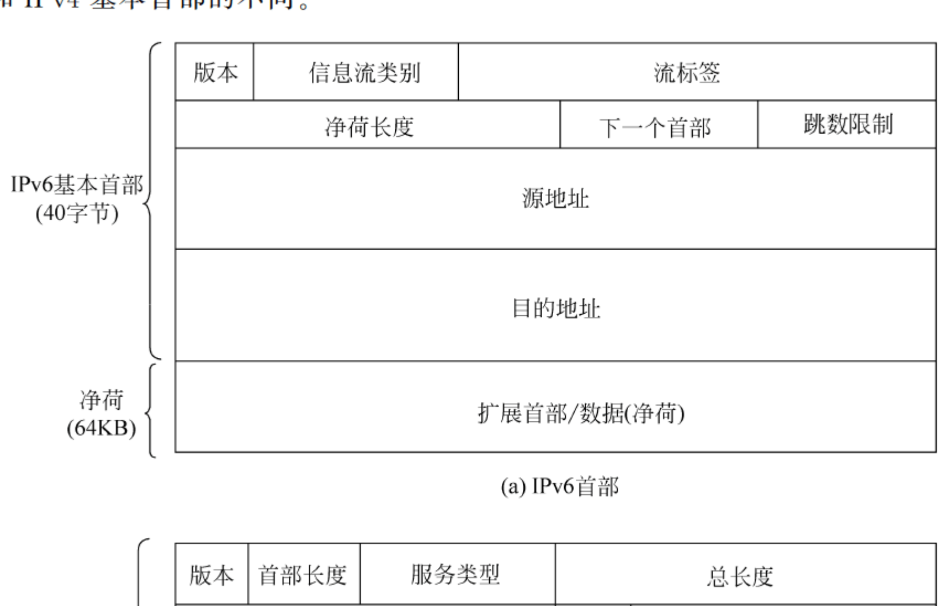
IPv4的以上种种不足表明确实需要根据目前Internet的规模和应用方式，提出一种新的既能体现IPv4分组交换灵活性，又能有效解决IPv4地址短缺、路由器转发处理复杂、路

由器流分类困难、信息资源安全机制先天不足等问题的网际协议，它就是 IPv6，也称下一代 IP。

10.2 IPv6 首部结构

10.2.1 IPv6 基本首部

IPv6 基本首部如图 10.1(a)所示，和图 10.1(b)的 IPv4 基本首部相比，去掉了和分片有关的字段及首部检验和字段，增加了流标签字段。下面详细介绍 IPv6 基本首部各字段的含义及和 IPv4 基本首部的不同。



(a) IPv6首部



(b) IPv4首部

图 10.1 IPv6 首部和 IPv4 首部

① 版本：4b，给出 IP 的版本号，IPv6 的版本号为 6，由于 IPv6 和 IPv4 的版本字段位于 IP 分组的同一位置，可用该字段值区分 IP 分组所属的 IP 版本。

② 信息流类别：8b，该字段给出 IP 分组对应的服务类别，其作用和 IPv4 的服务类型字段相同，在采用区分服务(Differentiated Services，DS)时，IPv6 的信息流类别字段和 IPv4 的服务类型字段值都是区分服务码点(Differentiated Services Code Point，DSCP)，用 DSCP 标识该 IP 分组的服务类别。

③ 流标签：20b，流是指一组具有相同的发送和接收进程的 IP 分组。分类服务有两大

类：一是区分服务(Differentiated Services,DiffServ)；二是综合服务(Integrated Services,IntServ)。区分服务定义若干服务类别,路由器为不同的服务类别设置不同的服务质量,当转发某个 IP 分组时,根据 IP 分组的服务类别字段值确定该 IP 分组所属的类别,并提供对应的服务质量。这种分类服务只能提供有限的服务类别,相同服务类别的 IP 分组具有相同的服务质量,当多个有着相同服务类别的 IP 分组在路由器中等待转发时,路由器按照先进先出的原则进行处理。综合服务是将属于特定会话的一组 IP 分组作为流,并为每一种流设置对应的服务质量。如两个 IP 电话之间的一次通话过程所涉及的 IP 分组就是一种流,路由器需要为该流预留带宽,以此保证两个 IP 电话之间的通话质量。区分服务是将信息流划分成有限的若干类,并为不同类别的信息流分配不同的服务质量,是宏观控制。综合服务是将信息流细分成流,并为每一种流设置相应的服务质量,是微观控制。路由器实施区分服务比较容易,但实施综合服务比较困难。实施综合服务首先需要确定 IP 分组所属的流,由于流是属于特定会话的一组 IP 分组,需要根据 IP 分组的源 IP 地址和目的 IP 地址、源端口号和目的端口号,甚至应用层 PDU 中的特定位置值来确定 IP 分组所属的流,这个过程比较复杂,会对路由器转发 IP 分组的速率产生严重影响。因此,一旦要求路由器实施综合服务,就需要牺牲转发速率。实际上,源终端在创建会话后,能够确定属于该会话的 IP 分组,因此,可以由源终端完成 IP 分组的流分类工作,并对属于特定流的 IP 分组分配唯一的流标签,路由器只需根据 IP 分组的源 IP 地址和流标签就可确定 IP 分组所属的流,并提供与该流对应的服务质量,这样就减少了路由器分类 IP 分组的处理负担,也符合目前尽量将处理功能转移到网络终端的趋势。因此,IPv6 首部中增加的流标签字段对路由器实施综合服务有莫大的帮助。

④ 净荷长度：16b,给出 IPv6 分组净荷的字节数。

⑤ 下一个首部：8b,IPv6 取消了可选项,增加了扩展首部,但扩展首部作为净荷的一部分出现在净荷字段中,这样,扩展首部的长度只受净荷字段长度的限制,而不像 IPv4,将可选项的总长度限制在 40B。当存在扩展首部时,用下一个首部给出扩展首部类型。当没有扩展首部时,该字段等同于 IPv4 的协议字段,用于指明净荷所属的协议。

⑥ 跳数限制：8b,给出 IP 分组允许经过的路由器数,IP 分组每经过一跳路由器,该字段值减 1,当该字段值减为 0 时,如果 IP 分组仍未到达目的终端,路由器将丢弃该 IP 分组,以此避免 IP 分组在网络中无休止地漂荡。IPv4 对应的是生存时间字段,它可以给出 IP 分组允许在网络中生存的时间,但实际上,路由器都将该字段作为跳数限制字段使用,因此,IPv6 使得该字段名符其实。

⑦ 源地址和目的地址：128b,源地址和目的地址字段的含义和 IPv4 相同,但 IPv6 的地址字段的长度是 128b,是 IPv4 的 4 倍,IPv6 彻底解决了 IPv4 面临的地址短缺问题。

IPv6 的首部长度是固定的,就是基本首部的长度,扩展首部属于净荷的一部分,因此,不需要首部长度字段。

对于 IPv4 分组,由于每经过一跳路由器,都会改变 TTL 字段值,需要重新计算首部检验和字段值,这将严重影响路由器的转发速率,而且,无论链路层还是传输层,都有差错控制功能,在目前通信链路的可靠性有所保证的前提下,在网络层重复差错控制功能的必要性不高,因此,IPv6 去掉首部检验和字段。

在 IPv4 中,当 IP 分组的长度超过输出链路的最大传输单元(Maximum Transmission Unit,MTU)时,由路由器负责将 IP 分组分片,因此,在 IP 首部中给出了和分片有关的字

段。在 IPv6 中,源终端通过协议获得源终端至目的终端传输路径所经过的链路的最小 MTU,并以此确定是否需要将 IP 分组分片,在需要分片的情况下,由源终端完成分片功能,因此,中间路由器是不涉及和分片有关的操作的,因此将和分片有关的字段放在分片扩展首部中。

10.2.2 IPv6 扩展首部

1. 扩展首部组织方式

IPv4 首部如果包含了可选项,中间经过的每一跳路由器都需要对可选项进行处理,增加了路由器的处理负担,降低了路由器转发 IPv4 分组的速率。IPv6 除了逐跳选项扩展首部外,中间路由器将扩展首部作为分组净荷对待,不对其作任何处理,以此简化路由器转发 IP 分组所进行的操作,提高路由器的转发速率。IPv6 目前定义的扩展首部有:逐跳选项、路由、分片、鉴别、封装安全净荷、目的端选项这 6 种,当 IP 分组包含多个扩展首部时,扩展首部按照以上顺序出现,上层协议数据单元(PDU)总是放在最后面。图 10.2 是上层协议数据单元(PDU)为 TCP 报文时,IPv6 分组的格式。

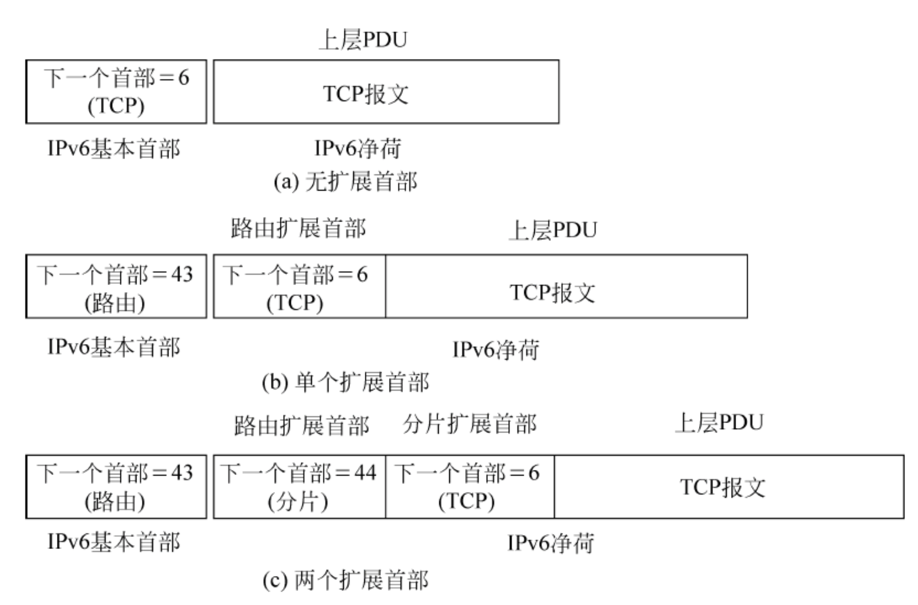


图 10.2 IPv6 基本首部、扩展首部和上层协议数据单元之间的关系

图 10.2(a)的 IPv6 分组没有扩展首部,净荷字段中只包含上层协议数据单元(TCP 报文),因此,基本首部中的下一个首部字段值给出上层协议类型 6,指明上层协议为 TCP。图 10.2(b)的IPv6 分组中包含单个扩展首部,净荷字段中首先出现的是路由扩展首部,而基本首部中的下一个首部字段值给出扩展首部的类型,扩展首部中的下一个首部字段值给出上层协议类型。图 10.2(c)的 IPv6 分组中包含两个扩展首部,依次在净荷字段中出现的是路由和分片扩展首部,基本首部中的下一个首部字段值给出第 1 个扩展首部的类型(路由),路由扩展首部中的下一个首部字段值给出第 2 个扩展首部的类型(分片),分片扩展首部中的下一个首部字段值给出上层协议类型(TCP)。当净荷字段中包含两个以上的扩展首部时,由前一个扩展首部中的下一个首部字段值给出下一个扩展首部的类型,最后一个扩展首部的下一个首部字段值给出上层协议类型。

2. 扩展首部应用实例

下面通过分片扩展首部的应用，说明 IPv6 简化路由器转发操作的过程。分片扩展首部格式如图 10.3 所示。它的各个字段的含义和 IPv4 首部中与分片有关的字段的含义相同，片偏移给出当前数据片在原始数据中的位置，标识符用来唯一标识分片数据后产生的数据片序列，接收端通过标识符鉴别出因为分片数据后产生的一组数据片。M 标志位用来标识最后一个数据片(M=0)。图 10.4 是一个互连网络结构图，链路上标出的数字是链路 MTU，对于 IPv4 分组，由路由器根据输出链路 MTU 和 IPv4 分组的总长确定是否对 IP 分组进行分片，并在需要分片的情况下，完成分片操作。对于 IPv6 分组，由源终端通过路径 MTU 发现协议找出源终端至目的终端传输路径所经过的链路的最小 MTU，该 MTU 称为路径 MTU，并由源终端完成分片操作，通过分片扩展首部给出各个数据片的片偏移及标识符。目的终端通过分片扩展首部中给出的信息，重新将各个数据片拼接成原始 IPv6 分组，整个操作过程如图 10.4 所示。

下一个首部	保留	片偏移	保留	M
标识符				

图 10.3 分片扩展首部

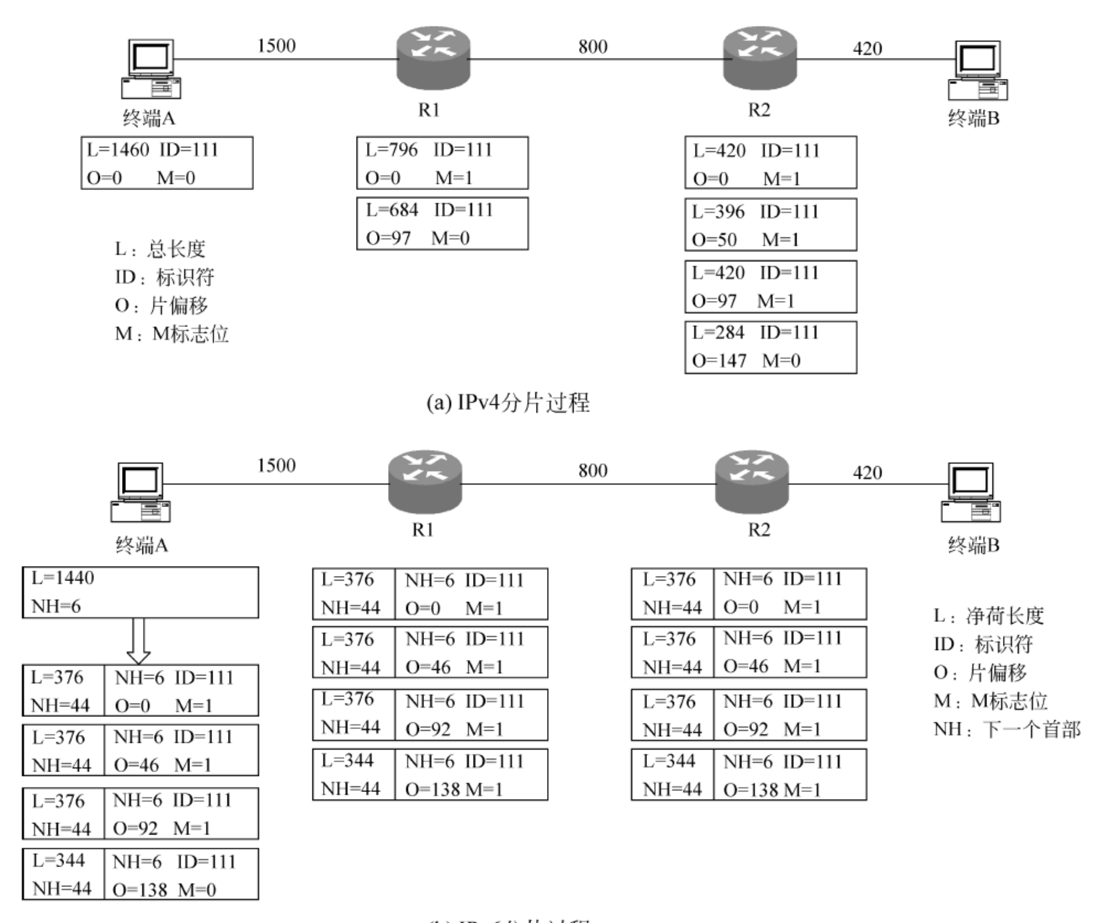


图 10.4 IPv4 和 IPv6 分片过程

IPv4 的分片操作过程已经在 5.2.4 节【例 5.4】中作了详细介绍，值得强调的是，IPv4 由路由器负责分片操作，而且可能由多个路由器对同一 IP 分组反复进行分片操作，如图 10.4(a)所示，这将严重影响路由器的转发速率，因此，在 IPv6 中，改由源终端完成分片操作。源终端首先通过路径 MTU 发现协议获取源终端至目的终端传输路径所经过的链路的最小 MTU(路径 MTU)，然后，对净荷进行分片，通常情况下，除最后一个数据片，其他数据片长度的分配原则是：必须是 8 的倍数，且加上 IPv6 首部和分片扩展首部后尽量接近路径 MTU。假定路径 MTU=M，净荷长度=L，将净荷分成 N 个数据片，则 $L+N\times48\leqslant M\times N$。48B 包括 40B IPv6 首部和 8B 分片扩展首部。在本例中，M=420B，L=1440B，根据 $1440+N\times48\leqslant420\times N$，得出 $N\geqslant1440/(420-48)=3.87$，$N$ 取满足上述等式的最小整数 4。前 3 个数据片长度应该是满足小于等于(420−48)B 且是 8 的倍数的最大值，这里是368B，加上 8B 的分片扩展首部后，得出净荷长度=376B，最后 1 个数据片的长度是 1440−3×368=336B，得出净荷长度=344B。4 个数据片的片偏移分别是 0、368/8=46、736/8=92、1104/8=138。值得说明的是，在每个会话的存在期间，源终端和目的终端之间都有大量 IP 分组传输，因此，源终端先通过路径 MTU 发现协议获取源终端至目的终端传输路径所经过的链路的最小 MTU(路径 MTU)是值得的，否则，对每一个 IP 分组都进行图 10.4(a)所示的分片操作会对路由器的转发速率造成巨大影响。

10.3 IPv6 地址结构

开发 IPv6 的主要原因是为了解决 IPv4 的地址短缺问题，因此，IPv6 的地址字段长度是 IPv4 的 4 倍：128b。有人计算过，2^{128}的 IPv6 地址空间可以为地球表面每平方米的面积提供 10.65×10^{23}个不同的 IPv6 地址，这么多的 IPv6 地址可以为地球上的每一粒沙子分配唯一的 IPv6 地址。如此巨大的地址空间，为使用 IPv6 地址提供了非常大的灵活性。

10.3.1 IPv6 地址表示方式

1. 基本表示方式

基本表示方式是将 128b 以 16 位为单位分段，每一段用 4 位十六进制数表示，各段用冒号分隔。下面是几个用基本表示方式表示的 IPv6 地址。

```
2001:0000:0000:0410:0000:0000:0001:45FF
0000:0000:0000:0000:0001:0765:0000:7627
```

2. 压缩表示方式

基本表示方式中可能出现很多 0，甚至可能整段都是 0，为了简化地址表示，可以将不必要的 0 去掉。不必要的 0 是指去掉后，不会错误理解段中 16 位二进制数的那些 0。如 0410 可以压缩成 410，但不能压缩成 41 或 041。上述用基本表示方式表示的 IPv6 地址可以压缩成如下表示方式。

```
2001:0:0:410:0:0:1:45FF
```

```
0:0:0:0:1:765:0:7627
```

用压缩表示方式表示的IPv6地址仍然可能出现相邻若干段都是0的情况，为了进一步缩短地址表示方式，可用一对冒号::表示连续的一串0，当然，一个IPv6地址只能出现一个::，这种用::表示连续的一串0的压缩表示方式就是0压缩表示方式，上述地址用0压缩表示方式表示如下。

```
2001::410:0:0:1:45FF
::1:765:0:7627
```

2001:0:0:410:0:0:1:45FF也可表示成2001:0:0:410::1:45FF，但不能表示成2001::410::1:45FF，因为后一种表示无法确定每一个::表示几个相邻的0。

【例10.1】 将下列用基本表示方式表示的IPv6地址用0压缩表示方式表示。

```
0000:0000:0000:0000:FE80:0000:0000:0000
0000:0001:1000:0000:0000:0000:0000:0000
0100:0000:0001:1000:0000:0000:0001:1000
```

【解析】 用0压缩表示方式表示如下。

```
::FE80:0: 0: 0
0:1:1000::
100:0:1:1000::1:1000
```

【例10.2】 将下述用0压缩表示方式表示的IPv6地址还原成基本表示方式。

```
::1:10:0:0
FE00:1000::
0:0:1::FE00
```

【解析】 上述用0压缩表示方式表示的IPv6地址还原成如下基本表示方式。

```
0000:0000:0000:0000:0001:0010:0000:0000
FE00:1000:0000:0000:0000:0000:0000:0000
0000:0000:0001:0000:0000:0000:0000:FE00
```

3. 特殊地址

1）内嵌IPv4地址的IPv6地址

这种地址是为了解决IPv4和IPv6共存时期配置不同版本的IP地址的终端之间通信问题而设置的，128b的地址中包含32b的IPv4地址，32b的IPv4地址仍然采用IPv4的地址表示方式，以8位为单位分段，每一段用对应的十进制值表示，段之间用点分隔，地址的其他部分采用IPv6的地址表示方式。以下是常用的2种内嵌IPv4地址的IPv6地址的表示方式。这两种地址的使用方式在后面章节中讨论。

```
0000:0000:0000:0000:0000:FFFF:192.167.12.16 或是::FFFF:192.167.12.16
0000:0000:0000:0000:FFFF:0000:192.167.12.16 或是::FFFF:0:192.167.12.16
```

2）环回地址

::1是IPv6的环回地址，等同于IPv4的127.X.X.X。

3）未确定地址

全0地址（表示成::）作为未确定地址，当某个没有分配有效IPv6地址的终端需要发送IPv6分组时，可用该地址作为IPv6分组的源地址。该地址不能作为IPv6分组的目的地址。

4. 地址前缀

IPv6采用无分类编址方式，将地址分成前缀部分和主机号部分，用前缀长度给出地址中表示前缀的二进制位数，用下述表示方式表示地址前缀。

IPv6地址/前缀长度

IPv6地址必须是用基本表示方式或0压缩表示方式表示的完整地址，前缀长度是一个0～128的整数，指出IPv6地址的高位中作为前缀的位数。下述是正确的前缀表示方式。

```
::FE80:0: 0: 0/68
::1:765:0:7627/60
2001:0000:0000:0410:0000:0000:0001:45FF/64
```

10.3.2 IPv6地址分类

IPv6地址分为单播、组播和任播这三种类型。

单播地址：唯一标识某个接口，以该种类型地址为目的地址的IP分组，到达目的地址标识的唯一的接口。

组播地址：标识一组接口，而且大部分情况下，这组接口分属于不同的结点（终端或路由器），以该种类型地址为目的地址的IP分组，到达所有由目的地址标识的接口。

任播地址：标识一组接口，而且大部分情况下，这组接口分属于不同的结点（终端或路由器），以该种类型地址为目的地址的IP分组，到达由目的地址标识的一组接口中的其中一个接口，该接口往往是这一组接口中和源终端距离最近的那个接口。

1. 单播地址

1）链路本地地址

链路不是物理线路，它指的是实现连接在同一网络的两个结点之间通信的传输网络，如以太网。链路本地地址指的是在同一传输网络内作用的IP地址，它的作用一是用于实现同一传输网络内两个结点之间的网络层通信，二是用于标识连接在同一传输网络上的接口，并用该IP地址解析接口的链路层地址。一旦某个接口被定义为IPv6接口，该接口自动生成链路本地地址。链路本地地址格式如图10.5所示。

10b	54b	64b
1111111010	0	接口标识符

图10.5 链路本地地址结构

链路本地地址的高64b是固定不变的，低64b是接口标识符。接口标识符用于在传输网络内唯一标识某个连接在该传输网络上的接口，它通常由接口的链路层地址导出。不同类型的传输网络导出接口标识符的过程不同，下面是通过以太网的MAC地址导出接口标

识符的过程。

48 位 MAC 地址由 24 位的公司标识符和 24 位的扩展标识符组成，公司标识符由 IEEE 负责分配。公司标识符最高字节的第 0 位是 I/G(单播地址/组地址)位，该位为 0，表明是单播地址，该位为 1，表明是组地址。第 1 位是 G/L(全局地址/本地地址)位，该位为 0，表明是全局地址，该位为 1，表明是本地地址。一般情况下，MAC 地址都是全局地址，G/L 位为 0。MAC 地址导出接口标识符的过程如图 10.6 所示，首先将 MAC 地址的 G/L 位置 1，然后在公司标识符和扩展标识符之间插入十六进制值为 FFFE 的 16 位二进制数。

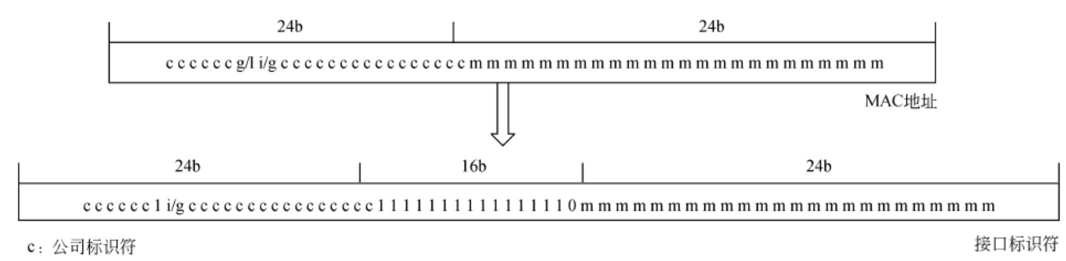
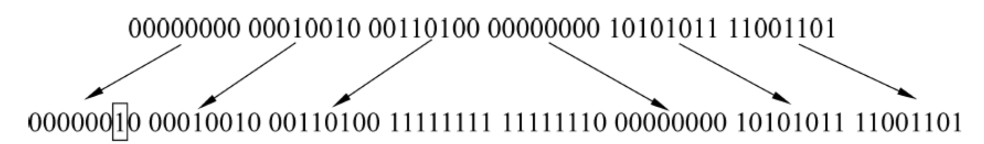


图 10.6 MAC 地址导出接口标识符过程

【例 10.3】 假定 MAC 地址为 0012:3400:ABCD，求接口标识符。

【解析】



接口标识符为 0212:34FF:FE00:ABCD。

【例 10.4】 假定 MAC 地址为 0012:3400:ABCD，求接口的链路本地地址。

【解析】 链路本地地址为 FE80:0000:0000:0000:0212:34FF:FE00:ABCD 或为 FE80::0212:34FF:FE00:ABCD。

2）站点本地地址

站点本地地址类似于 IPv4 的本地地址(或称私有地址)，它不是全球地址，只能在内部网络内使用。和链路本地地址不同，它可以用于标识内部网络内连接在不同子网上的接口。因此，除了接口标识符字段外，还有子网标识符字段，用子网标识符字段标识接口所连接的子网。站点本地地址不能自动生成，需要配置，在手工配置站点本地地址时，接口标识符可以和链路本地地址的接口标识符一样，通过接口的链路层地址导出，也可手工配置一个子网内唯一的标识符作为接口标识符。站点本地地址格式如图 10.7 所示。

10b	38b	16b	64b
1111111011	0	子网标识符	接口标识符

图 10.7 站点本地地址结构

3）可聚合全球单播地址

可聚合全球单播地址格式如图 10.8 所示。

48b	16b	64b
全球路由前缀	子网标识符	接口标识符

图 10.8 可聚合全球单播地址结构

图 10.8 将地址分成三级，它们分别是全球路由前缀、子网标识符和接口标识符，全球路由前缀用于 Internet 主干网中路由器为 IPv6 分组选择传输路径，因此，分配全球路由前缀时，要求尽可能将高 N 位相同的全球路由前缀分配给同一物理区域，如将高 5 位相同的全球路由前缀分配给亚洲，而将高 8 位相同的全球路由前缀分配给中国，当然，高 8 位中的最高 5 位和分配给亚洲的全球路由前缀的高 5 位相同，以此最大可能聚合路由项。应该说，除了已经分配的 IPv6 地址空间外，其余的地址空间都可分配作为可聚合全球单播地址，但目前已经指定作为可聚合全球单播地址的是最高 3 位为 001 的 IPv6 地址空间。子网标识符用于标识划分某个公司或组织的内部网络所产生的子网。接口标识符用来确定连接在某个子网上的接口。需要说明的是，上述地址结构只是在全球范围内分配 IPv6 地址时有用，在转发 IPv6 分组时，路由项中的地址只有两部分：网络前缀和主机号，没有图 10.8 所示的地址结构。在全球范围内分配 IPv6 地址时采用图 10.8 所示的地址结构和尽可能将高 N 位相同的全球路由前缀分配给同一物理区域的目的是为了尽可能地聚合路由项，减少路由表中路由项的数目，提高转发速率。图 10.9 给出了尽可能聚合路由项的全球路由前缀分配过程。

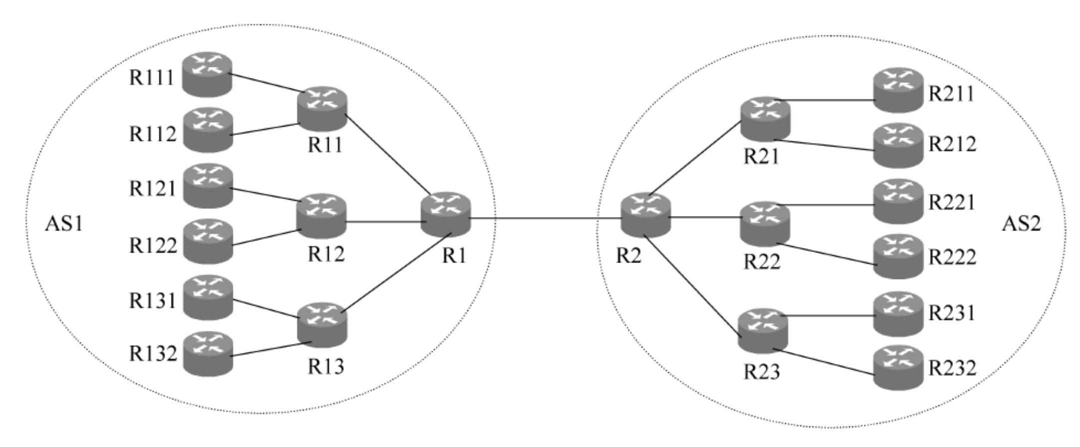


图 10.9 尽可能聚合路由项的全球路由前缀分配过程

对于图 10.9 所示的网络结构，为 AS1 和 AS2 分别分配高 5 位相同的全球路由前缀，如 00100 和 00101。为 AS1 中路由器 R11、路由器 R12 和路由器 R13 连接的三个分枝分别分配高 8 位相同的全球路由前缀，如 00100000、00100001 和 00100010。为 R111 等路由器连接的分枝分别分配高 12 位相同的全球路由前缀，如 001000000000。其他路由器连接的分枝以此类推，可以得出表 10.1 所示的地址分配结构。

表 10.1 地址结构

路由器	全球路由前缀	子网标识符	接口标识符
R111	00100 000 0000	X	X:X:X:X
R112	00100 000 0001	X	X:X:X:X
R121	00100 001 0000	X	X:X:X:X

续表

路由器	全球路由前缀	子网标识符	接口标识符
R122	00100 001 0001	X	X:X:X:X
R131	00100 010 0000	X	X:X:X:X
R132	00100 010 0001	X	X:X:X:X
R211	00101 000 0000	X	X:X:X:X
R212	00101 000 0001	X	X:X:X:X
R221	00101 001 0000	X	X:X:X:X
R222	00101 001 0001	X	X:X:X:X
R231	00101 010 0000	X	X:X:X:X
R232	00101 010 0001	X	X:X:X:X

根据表 10.1 给出的地址结构，可以得出如表 10.2 所示的路由器 R1 用于指明通往图 10.9 中所有网络的传输路径的路由项。

表 10.2 路由器 R1 路由表

目的网络	下一跳	备 注
2800::/5	R2	指向 AS2 的路由项
2000::/8	R11	指向 R11 连接的分枝的路由项
2100::/8	R12	指向 R12 连接的分枝的路由项
2200::/8	R13	指向 R13 连接的分枝的路由项

从表 10.1 中可以看出，由于为每一个分枝所连接的网络分配了高 N 位相同的全球路由前缀，只需一项路由项就可以指出通往某个分枝所连接的所有网络的传输路径。

2. 组播地址

组播地址格式如图 10.10 所示，高 8 位固定为十六进制值 FF，4 位标志位中的前 3 位固定为 0，最后 1 位如果为 0，表示是由网络号码指派管理局(Internet Assigned Numbers Authority，IANA)分配的永久组播地址，这些组播地址有特定用途，因而也被称为著名组播地址。最后位如果为 1，表示是非永久分配的组播地址(临时的组播地址)。范围字段中正常使用的值如下：

2：链路本地范围；

5：站点本地范围；

8：组织本地范围；

E：全球范围。

链路本地范围是指组播只能在单个传输网络范围内进行。站点本地范围是指组播在有多个传输网络组成的站点网络内进行。组织本地范围是指组播在有多个站点网络组成，但由同一组织管辖的网络内进行。全球范围指在 Internet 中组播。

8b	4b	4b	80b	32b
11111111	标志	范围	0	组标识符

图 10.10 组播地址结构

IANA 分配的常用著名组播地址有：

FF02::1 链路本地范围内所有结点；

FF02::2 链路本地范围内所有路由器；

FF05::2 站点本地范围内所有路由器；

FF02::9 链路本地范围内所有运行 RIP 的路由器。

3. 任播地址

没有为任播地址分配单独的地址格式，在单播地址空间中分配任播地址，如果为某个接口分配了任播地址，必须在分配地址时说明。目前只有路由器接口允许分配任播地址，本教材不对任播地址的应用方式进行讨论。

10.4 IPv6 操作过程

实现图 10.11 中的终端 A 至终端 B 的数据传输，必须完成两方面的操作，一是网络层必须完成如下操作：

- 终端配置全球 IPv6 地址；
- 终端配置默认路由器地址；
- 路由器建立路由表。

二是连接终端和路由器及互连路由器的传输网络必须完成 IPv6 over X(X 指不同类型的传输网络)操作，本节讨论 IPv6 的网络层操作过程，下一节讨论 IPv6 over 以太网操作过程。

在 IPv4 网络中，路由器接口地址手工配置，终端接口的 IPv4 地址和默认路由器地址可以手工配置，也可通过动态主机配置协议(Dynamic Host Configuration Protocol，DHCP)自动获取。路由器中的路由表通过路由协议动态建立。

在 IPv6 网络中，可以为路由器接口配置多种类型的地址，一种是全球地址，需要手工配置；另一种是链路本地地址，在指定某个接口为 IPv6 接口后，由路由器自动生成。终端接口也有两种类型的接口地址，一种是全球地址，用于向其他网络中的终端传输数据；另一种是只在终端接口所连接的传输网络内作用的链路本地地址，在指定终端接口为 IPv6 接口后，由终端自动生成。终端接口的全球地址和默认路由器地址与 IPv4 网络一样，可以手工配置，也可以通过 DHCP 自动获取。如果手工配置，配置人员必须了解终端所连接子网的拓扑结构和路由器配置信息。如果通过 DHCP 自动获取，必须管理、同步 DHCP 服务器内容。由于 IPv6 可能被未来家电用于数据传输，而人们对家电总是希望即插即用，不愿意在对家电进行配置或向某个管理人员注册后启用家电。为此，IPv6 提供了邻站发现(Neighbor Discovery，ND)协议，以此来解决 IPv6 终端的即插即用问题。

10.4.1 邻站发现协议

1. 终端获取全球地址和默认路由器地址过程

终端将接口定义为 IPv6 接口后，自动为接口生成链路本地地址，在图 10.11 中，假定终

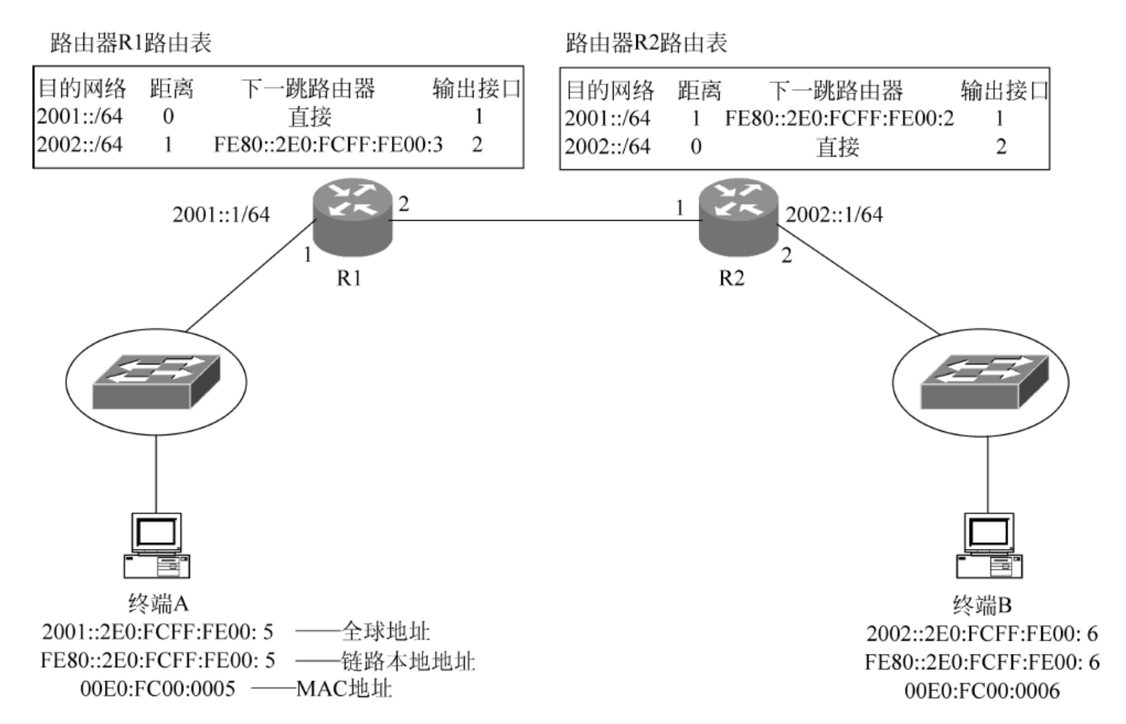


图 10.11　IPv6 网络结构

端 A 和终端 B 的 MAC 地址分别为 00E0:FC00:0005 和 00E0:FC00:0006,终端 A 和终端 B 分别生成链路本地地址 FE80::2E0:FCFF:FE00:5 和 FE80::2E0:FCFF:FE00:6。同样,根据路由器 R1、路由器 R2 的接口 1 和接口 2 的 MAC 地址分别求出如表 10.3 所示的链路本地地址。

表 10.3　路由器各个接口的链路本地地址

路由器接口	MAC 地址	链路本地地址
路由器 R1 接口 1	00E0:FC00:0001	FE80::2E0:FCFF:FE00:1
路由器 R1 接口 2	00E0:FC00:0002	FE80::2E0:FCFF:FE00:2
路由器 R2 接口 1	00E0:FC00:0003	FE80::2E0:FCFF:FE00:3
路由器 R2 接口 2	00E0:FC00:0004	FE80::2E0:FCFF:FE00:4

终端 A 和终端 B 分别求出链路本地地址后,需要求出接口的全球地址和默认路由器地址,由于终端和默认路由器连接在同一个传输网络,具有相同的网络前缀,因此,终端只要得到默认路由器的网络前缀和通过接口的链路层地址导出的接口标识符就可得出全球地址。由于接口标识符为 64 位,因此,网络前缀也必须是 64 位,这样才能组合出 128 位的全球地址。现在的问题是终端如何获取默认路由器地址和网络前缀。

IPv6 路由器定期通过各个接口组播路由器通告,该通告的源地址是发送接口的链路本地地址,目的地址是表明接收方是链路中所有结点的著名组播地址 FF02::1,通告中给出为接口配置的全球地址的网络前缀、前缀长度及路由器生存时间等参数。当终端接收到某个路由器通告,该通告的源地址就是路由器连接终端所在网络的接口的地址,就是默认路由器地址,通告中给出的网络前缀和前缀长度即是终端在网络的网络前缀,当该网络前缀的长度为 64 位时,终端将其和通过终端接口链路层地址导出的接口标识符组合在一起,构成 128

位的终端全球地址。为了将这种全球地址获取方式和通过 DHCP 服务器的自动获取方式相区别,称这种地址获取方式为无状态地址自动配置,而称通过 DHCP 服务器获取地址的方式为有状态地址自动配置。由于路由器是定期发送路由器通告,因此当某个终端启动后,可能需要等待一段时间才能接收到路由器通告。如果终端希望立即接收到路由器通告,终端可以向路由器发送路由器请求,该路由器请求的源地址是终端接口的链路本地地址,目的地址是表明接收方是链路中所有路由器的著名组播地址 FF02::2。当路由器接收到路由器请求,立即组播一个路由器通告。图 10.12 给出了终端获取全球地址及默认路由器地址的过程。

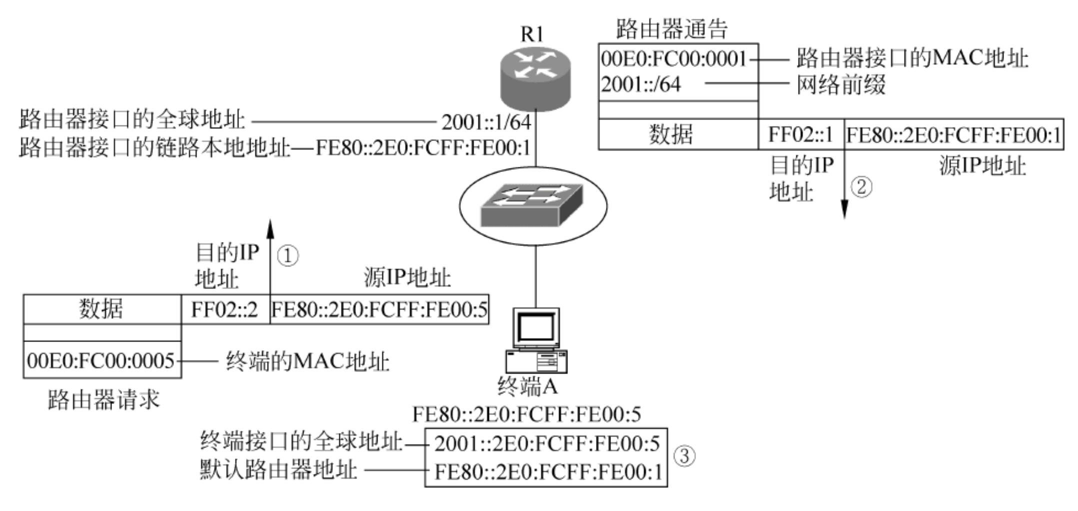


图 10.12 终端获取网络前缀和默认路由器地址过程

从图 10.12 中可以看出,无论是终端发送的路由器请求,还是路由器发送的路由器通告,都给出了发送接口的链路层地址(这里是以太网的 MAC 地址),这主要因为 IPv6 分组必须封装在 MAC 帧的数据字段中,才能通过传输网络传输给下一跳结点,因此,在通过传输网络传输 IPv6 分组前,必须先获取下一跳结点的 MAC 地址,在路由器请求和通告中给出发送接口的链路层地址就是为了这一目的。IPv4 over 以太网通过 ARP 实现地址解析,即根据下一跳结点的 IPv4 地址获取下一跳结点的 MAC 地址,IPv6 通过邻站发现协议解决这一问题,下一节 IPv6 over 以太网将详细讨论 IPv6 的地址解析过程。

2. 重复地址检测

无论是链路本地地址,还是通过无状态地址自动配置方式得出的全球地址,其唯一性都依赖于接口标识符的唯一性,由于不同网络的网络前缀是不同的,因此,只要保证同一网络内不存在相同的接口标识符,就可保证地址的唯一性。重复地址检测(Duplicate Address Detection,DAD)就是用来确定网络中是否存在另一个和某个接口有着相同的接口标识符的接口。

当结点的某个接口自动生成了 IPv6 地址(链路本地地址或全球地址),结点通过该接口发送邻站请求来确定该地址的唯一性,该邻站请求的接收方应该是可能具有相同接口标识符的接口,为此,对任何进行重复检测的单播地址,都定义了用于指定可能具有相同接口标识符的接口集合的组播地址,该组播地址的网络前缀为 FF02::1:FF00:0/104,低 24 位为单播地址的低 24 位,实际上就是接口标识符的低 24 位。这就意味着链路中所有接口标识

符低 24 位相同的接口组成一个组播组,以该组播地址为目的地址的 IP 分组被该组播组中的所有接口接收。某个接口的地址在通过重复地址检测前属于试验地址,不能正常使用,因此,某个源结点为确定接口地址唯一性而发送的邻站请求,其源地址为未确定地址::(全0),目的地址是根据需要进行重复检测的接口地址的低 24 位导出的组播地址。邻站请求中的目标地址字段给出需要重复检测的单播地址,即试验地址。当属于由目的地址指定的组播组的接口(接口标识符低 24 位和需要重复地址检测的单播地址的低 24 位相同的接口)接收到邻站请求,接收到邻站请求的结点(目的结点)用接收邻站请求的接口的接口地址和邻站请求中包含的试验地址比较,如果相同,且该接口的接口地址也是试验地址,该接口将放弃使用该试验地址。如果该接口的接口地址是正常使用的地址(非试验地址),目的结点向源结点发送邻站通告,该通告的源地址是接收邻站请求的接口正常使用的接口地址,目的地址是表明接收方是链路中所有结点的组播地址 FF02::1,通告中目标地址字段给出对应的邻站请求中的目标地址字段值和该接口的链路层地址。如果目的结点接收到邻站请求的接口的接口地址和邻站请求中包含的试验地址不同,目的结点不作任何处理。当源结点发送邻站请求后,如果接收到邻站通告,且通告中包含的目标地址字段值和接口的试验地址相同,源结点将放弃使用该接口地址。如果源结点发送邻站请求后,在规定时间内一直没有接收到对应的邻站通告,确定链路中不存在和其接口标识符相同的其他接口,将该接口地址作为正常使用的地址。整个过程如图 10.13 所示。

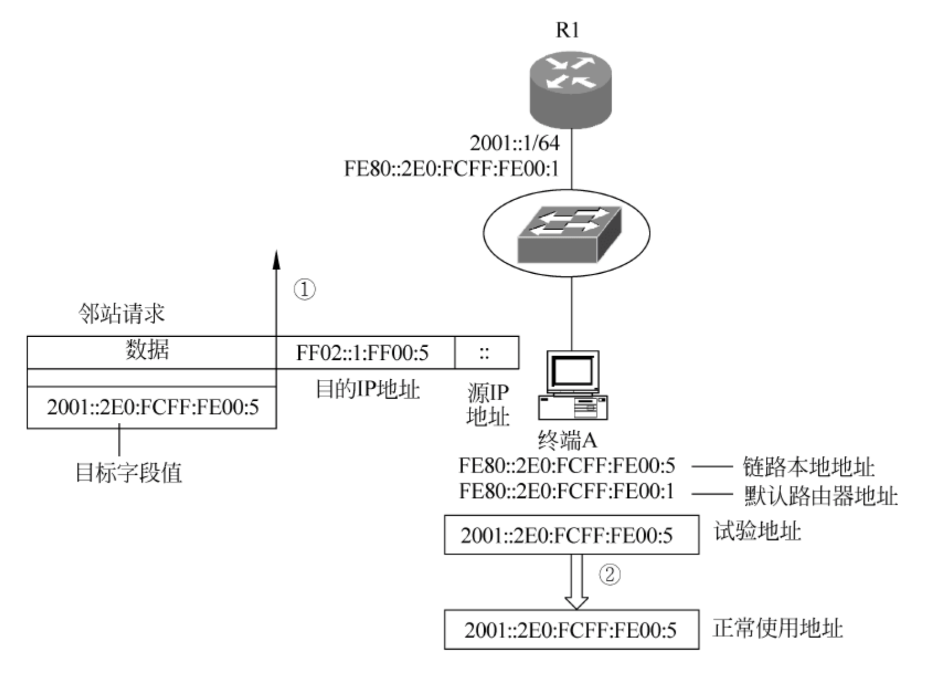


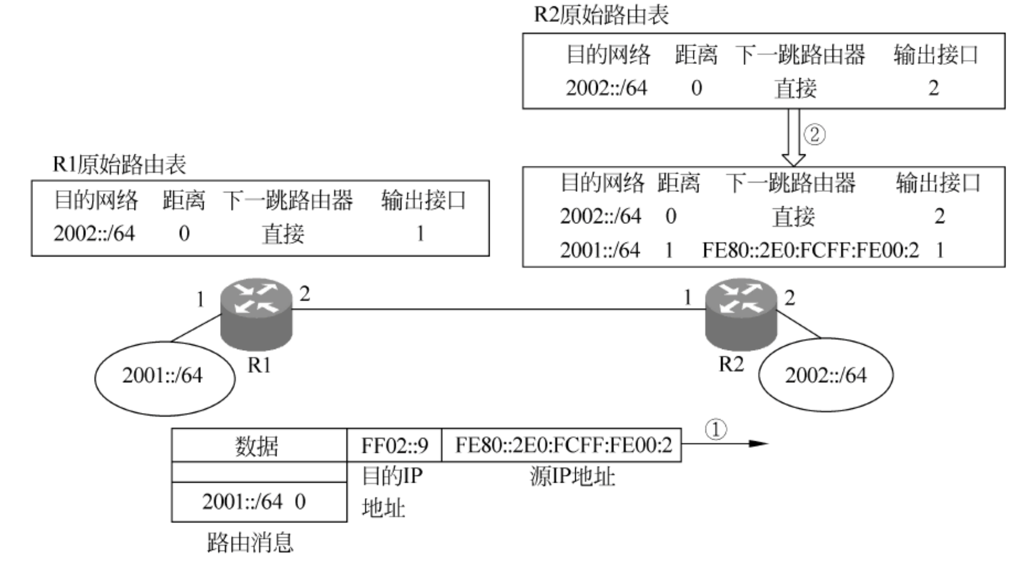
图 10.13 重复地址检测过程

10.4.2 路由器建立路由表过程

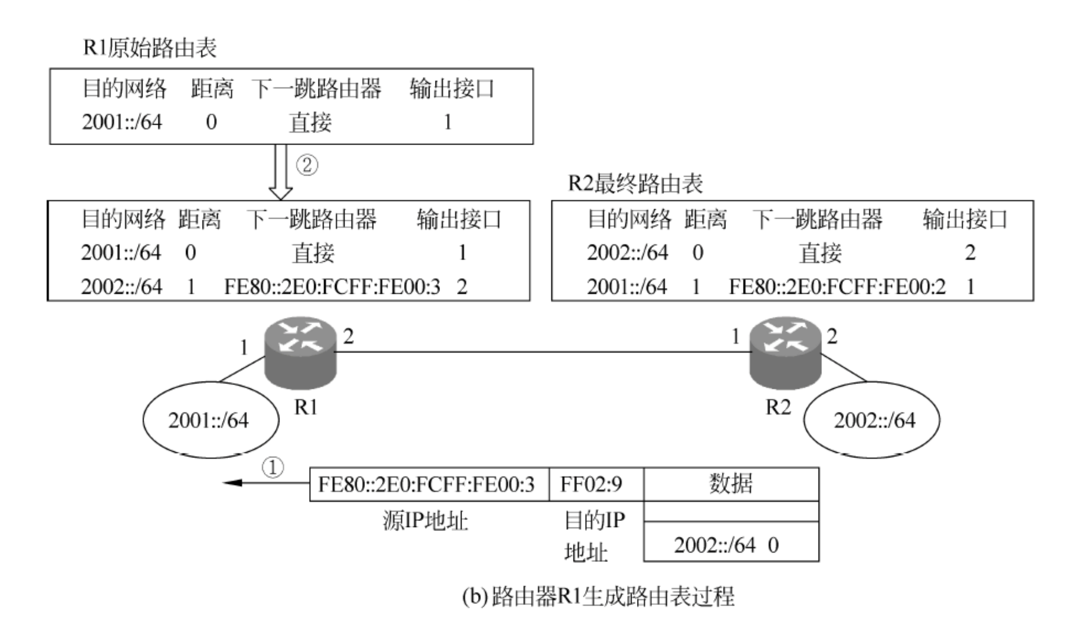
IPv6 中路由器通过路由协议建立路由表的过程和 IPv4 基本相同,只是路由项中的目的网络用 IPv6 地址的网络前缀表示方式表示。封装路由消息的 IP 分组的源地址是发送该路由消息的接口的链路本地地址,目的地址是表示链路本地范围内所有运行 RIP 的路由器的著名组

播地址：FF02::9。因此，路由表中下一跳路由器地址也是下一跳路由器对应接口的链路本地地址。下面通过用下一代 RIP(RIP Next Generation，RIPng)建立图 10.11 所示 IPv6 网络结构中路由器 R1 和路由器 R2 的路由表为例，讨论在 IPv6 网络中路由器建立路由表的过程。

当路由器 R1 的接口 1 和路由器 R2 的接口 2 配置了全球地址和网络前缀后，路由器 R1、路由器 R2 自动生成图 10.14 所示的初始路由表。然后路由器 R1 和路由器 R2 周期性地向对方发送包含路由表中路由项的路由消息。图 10.14(a)是路由器 R1 向路由器 R2 发送路由消息的过程，当路由器 R2 接收到路由器 R1 发送的路由消息，进行 6.3.2 节中讨论的 RIP 路由消息处理流程，在路由表中增添用于指明通往网络 2001::/64 的传输路径的路由项。同样，路由器 R2 也向路由器 R1 发送路由消息，使得路由器 R1 也得出指明通往网络 2002::/64 的传输路径的路由项，整个过程如图 10.14(b)所示。



(a) 路由器R2生成路由表过程



(b) 路由器R1生成路由表过程

图 10.14 路由器建立路由表过程

10.5 IPv6 over 以太网

10.5.1 地址解析过程

当图 10.11 中终端 A 想给终端 B 发送数据时,终端 A 构建一个以终端 A 的 IPv6 地址 2001::2E0:FCFF:FE00:5 为源地址,以终端 B 的 IPv6 地址 2002::2E0:FCFF:FE00:6 为目的地址的 IPv6 分组。终端 A 在开始发送该 IPv6 分组前,先检索路由表。根据图 10.11 所示的配置,终端 A 的路由表中存在如表 10.4 所示的 2 项路由项。和 IPv4 相同,终端的路由表内容通过手工配置和邻站发现协议获得,不是通过路由协议获得。

表 10.4 终端 A 建立的路由表

目 的 网 络	下一跳路由器
2001::/64	本地连接
::/0	FE80::2E0:FCFF:FE00:1

第 1 项指明终端 A 所连接的网络的网络前缀,第 2 项指明默认路由。和 IPv4 一样,终端 A 首先确定 IPv6 分组的目的终端是否和源终端连接在同一个网络(在 IPv6 网络,称为 on-link),这个过程需要比较目的地址的网络前缀和终端 A 所连接的网络的网络前缀。由于目的地址的网络前缀 2002::/64 和终端 A 所连接网络的网络前缀 2001::/64 不同,确定源终端和目的终端不在同一个网络(在 IPv6 网络,称为 off-link),终端 A 选择将该 IPv6 分组发送给默认路由器。在获取默认路由器的 IPv6 地址后,在将 IPv6 分组封装成经过以太网传输的 MAC 帧前,需要根据默认路由器的 IPv6 地址解析出 MAC 地址。这一过程称为地址解析过程,对应的 IPv6 地址称为解析地址。在前一节讨论终端获取网络前缀和默认路由器地址的过程(无状态地址自动配置过程)中已经讲到,路由器在链路本地范围内组播的路由器通告不仅包含网络前缀,而且还包含路由器连接该链路的接口的链路层地址,如果链路是以太网,接口的链路层地址就是 MAC 地址。因此,终端在完成获取网络前缀和默认路由器地址的过程后,不仅建立如表 10.4 所示的 2 项路由项,而且还建立如表 10.5 所示的邻站缓存,邻站缓存中的每一项给出邻站的 IPv6 地址和对应的链路层地址。如果在邻站缓存中找到默认路由器的 IPv6 地址对应的项,终端 A 可以立即通过该项给出的 MAC 地址封装 MAC 帧。否则,需要通过地址解析过程来获取默认路由器的 MAC 地址。和 IPv4 的 ARP 缓存相同,邻站缓存中的每一项都有寿命,如果在寿命内没有接收到用于确认 IPv6 地址和对应的链路层地址之间关联的信息,该项将因为过时而不再有效。这种情况下,终端也将通过地址解析过程获取和某个 IPv6 地址关联的链路层地址。

表 10.5 终端 A 邻站缓存

邻站 IPv6 地址	邻站链路层地址
FE80::2E0:FCFF:FE00:1	00E0:FC00:0001

地址解析过程首先由需要解析地址的终端发送邻站请求,邻站请求的源地址是发送该

邻站请求的接口的 IPv6 地址，由于每一个接口有多个 IPv6 地址，如终端 A 连接链路的接口有链路本地地址和全球地址，选择作为邻站请求的源地址的原则是选择最有可能被邻站用来解析接口的链路层地址的 IPv6 地址。由于终端 A 用全球地址作为发送给终端 B 的 IPv6 分组的源地址，那么，终端 B 回送给终端 A 的 IPv6 分组必定以终端 A 的全球地址作为目的地址。当路由器 R1 通过以太网传输终端 B 回送给终端 A 的 IPv6 分组时，需要通过该 IPv6 分组的目的地址解析终端 A 的链路层地址，因此，在这次数据传输过程中，路由器 R1 最有可能用来解析终端 A 的链路层地址的接口地址是全球地址。因此，终端 A 用接口的全球地址作为邻站请求的源地址。邻站请求的目的地址是组播地址，组播组标识符是解析地址的低 24 位，表示接收方是接口地址低 24 位等于组播组标识符的接口。邻站请求包含解析地址和发送邻站请求的接口的链路层地址。所有接口地址的低 24 位和解析地址的低 24 位相同的接口都接收该邻站请求，目的结点首先在邻站缓存中检索邻站请求源地址对应的项，如果找到对应项且对应项给出的链路层地址和邻站请求中给出的链路层地址相同，更新寿命定时器，否则，在邻站缓存中记录下源地址和链路层地址之间的关联。如果发现接收邻站请求的接口具有和解析地址相同的接口地址，目的结点回送邻站通告，通告中给出解析地址和解析出的链路层地址。终端 A 解析出默认路由器的链路层地址的过程如图 10.15 所示。

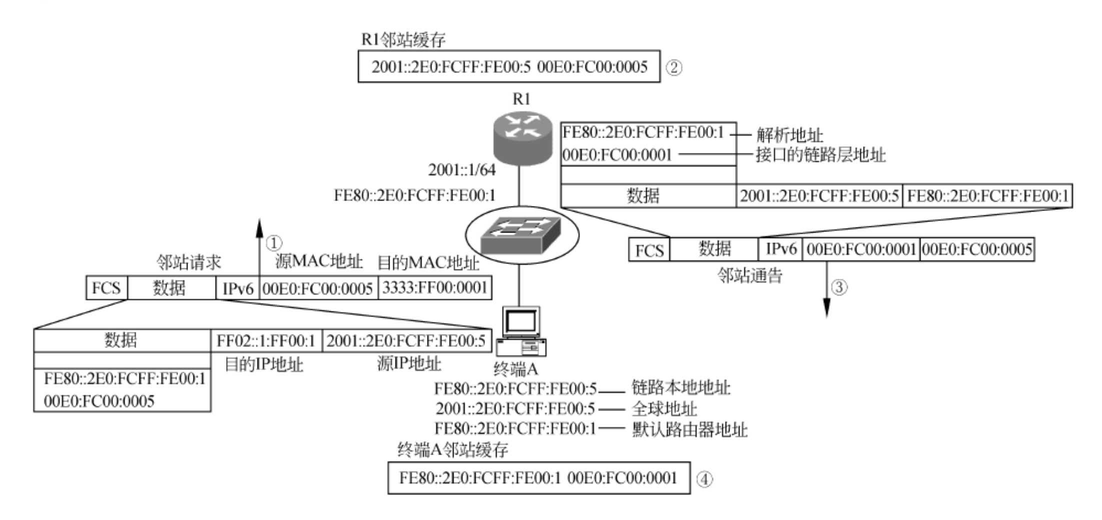


图 10.15 终端 A 解析出默认路由器的链路层地址的过程

上一节讨论重复地址检测时用到的也是邻站请求和邻站通告，这一节同样用邻站请求和邻站通告完成地址解析，目的结点必须区分出接收到的邻站请求是用于完成重复地址检测的邻站请求，还是用于完成地址解析的邻站请求。目的结点通过接收到的邻站请求的源地址区分出两种不同用途的邻站请求。由于通过重复地址检测前，分配给接口的地址是试验地址，不能正常使用，因此，邻站请求的源地址是未确定地址——::。而进行地址解析时，邻站请求的源地址是发送接口的正常使用地址。不同用途下邻站请求包含的目标地址字段值也不同，重复地址检测时发送的邻站请求中的目标地址字段给出用于进行重复检测的试验地址，而地址解析时发送的邻站请求中的目标地址字段给出用于解析出邻站链路层地址的邻站 IPv6 地址。

10.5.2 IPv6 组播地址和 MAC 组地址之间的关系

终端 A 组播的邻站请求被封装成 MAC 帧后，才能通过以太网进行传输，IPv6 分组封装成 MAC 帧的过程和 IPv4 相同，只是类型字段给出的十六进制值是 86DD，表明数据字段中数据的类型是 IPv6 分组。由于邻站请求是组播分组，目的 MAC 地址是根据 IPv6 组播地址转换成的 MAC 组地址。IPv6 组播地址转换成 MAC 组地址的过程如图 10.16 所示，MAC 组地址的高 16 位固定为 3333，低 32 位是 IPv6 组播地址的低 32 位。

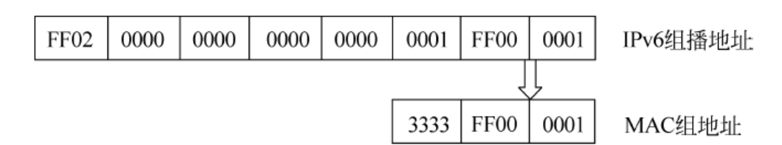


图 10.16 IPv6 组播地址转换成 MAC 组地址的过程

10.5.3 IPv6 分组传输过程

终端 A 解析出默认路由器 R1 的 MAC 地址后，将传输给终端 B 的 IPv6 分组封装成 MAC 帧，并通过以太网将该 MAC 帧传输给路由器 R1。路由器 R1 从接收到的 MAC 帧中分离出 IPv6 分组，用 IPv6 分组的目的地址检索路由表，找到下一跳路由器。同样用下一跳路由器的 IPv6 地址解析出下一跳路由器的 MAC 地址，再将 IPv6 分组封装成 MAC 帧，经过以太网将该 MAC 帧传输给路由器 R2。经过逐跳转发，最终到达终端 B，整个传输过程如图 10.17 所示。

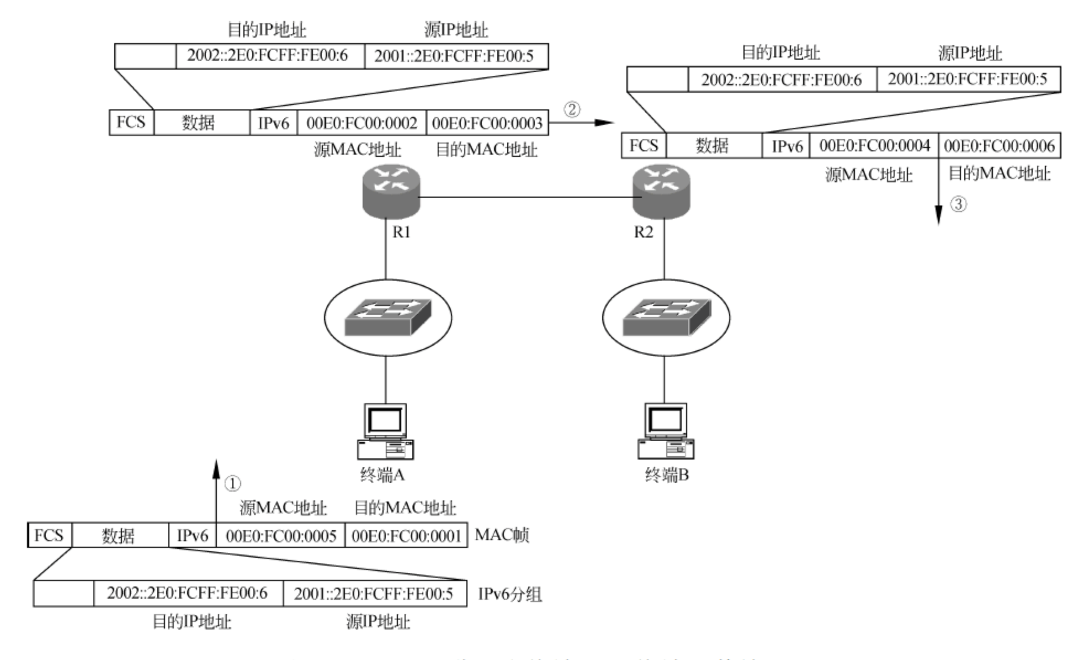


图 10.17 IPv6 分组由终端 A 至终端 B 传输过程

讨论 IPv4 over 以太网时讲过，IPv4 over 以太网涉及三方面内容：地址解析、IPv4 分组封装和 MAC 帧传输。IPv6 over 以太网同样涉及这三方面内容，除了地址解析过程，其余

两方面内容和 IPv4 完全相同。IPv4 的地址解析过程通过 ARP 实现，ARP 报文被直接封装成 MAC 帧，因此，ARP 只能实现类似以太网的广播型网络的地址解析过程，这就意味着 IPv4 对不同的传输网络，采用不同的地址解析协议。而邻站发现协议以 IPv6 分组格式传输协议报文，和传输网络无关，因此，IPv6 地址解析协议独立于传输网络，不同传输网络均可用邻站发现协议实现地址解析过程。更重要的是由于通过 IPv6 的鉴别和封装安全净荷扩展首部，可以对源终端进行鉴别，避免了其他终端冒用源终端的情况发生，因此，也不会出现类似 ARP 欺骗攻击这样的问题。ARP 欺骗攻击是指某个终端通过发送 ARP 请求，把别的终端的 IPv4 地址和自己的 MAC 地址绑定在一起，以此实现以窃取发送给别的终端的 IPv4 分组为目的的攻击手段。

10.6 IPv6 网络和 IPv4 网络互连

在 10.1 节中讨论网络互连时已经讲到，必须通过一种高于传输网络且独立于传输网络的协议来解决互连问题。因此，如果真正要求实现 IPv4 网络和 IPv6 网络互连，仿照通过 IP 实现不同类型的传输网络互连的方式，必须设计出一种高于 IPv4 和 IPv6 且独立于 IPv4 和 IPv6 的协议，这种协议能够对 IPv4 网络和 IPv6 网络中的终端分配统一的、独立于 IPv4 和 IPv6 的协议地址，因而可以在这一层的协议数据单元(PDU)中对位于 IPv6 或 IPv4 网络的源和目的终端给出统一的协议地址。而实现这一层协议的设备应该是某种网关设备，源终端至目的终端的传输路径由一系列这样的网关组成，而互连网关的网络是 IPv6 或 IPv4 网络，该协议数据单元通过 X over IPv4 或 X over IPv6(X 是指独立于 IPv6 和 IPv4 的上一层协议)技术实现相邻网关之间的传输。如果 IPv6 和 IPv4 网络也像不同类型的传输网络那样独立发展，实现 IPv4 网络和 IPv6 网络的互连必须走上述道路。但事实是，IPv6 网络和 IPv4 网络共存是暂时的，最终是 IPv6 网络取代 IPv4 网络，因此，实现 IPv4 网络和 IPv6 网络互连的需求也是暂时的。因而只能采用一些简单的方法来解决共存时期的通信问题，而不会像用 IP 实现不同类型的传输网络互连那样开发出一整套的协议和设备来实现 IPv4 网络和 IPv6 网络互连。

10.6.1 双协议栈技术

IPv4 和 IPv6 虽然互不兼容，各自有着独立的编址空间，但它们为传输层提供的服务是相同的，而且 IPv4 over X 技术和 IPv6 over X(X 指各种类型的传输网络)又十分相似，因此，人们开始生产同时支持 IPv4 和 IPv6 的路由器，这种路由器称为双协议栈路由器。由这种路由器构成的网络中，允许同时存在 IPv4 和 IPv6 终端，当然，IPv4 终端只能和另一个 IPv4 终端通信，IPv6 也同样。如果某个终端希望既能和 IPv4 又能和 IPv6 终端通信，这个终端也必须支持双协议栈。双协议栈体系结构如图 10.18 所示。



图 10.18 双协议栈结构

图 10.19 是采用双协议栈路由器的网络结构，路由器一旦采用双协议栈，它同时运行 IPv4 协议系列和 IPv6 协议系列，必须将所有接口定义为 IPv4 和 IPv6 接口，为接口分配 IP

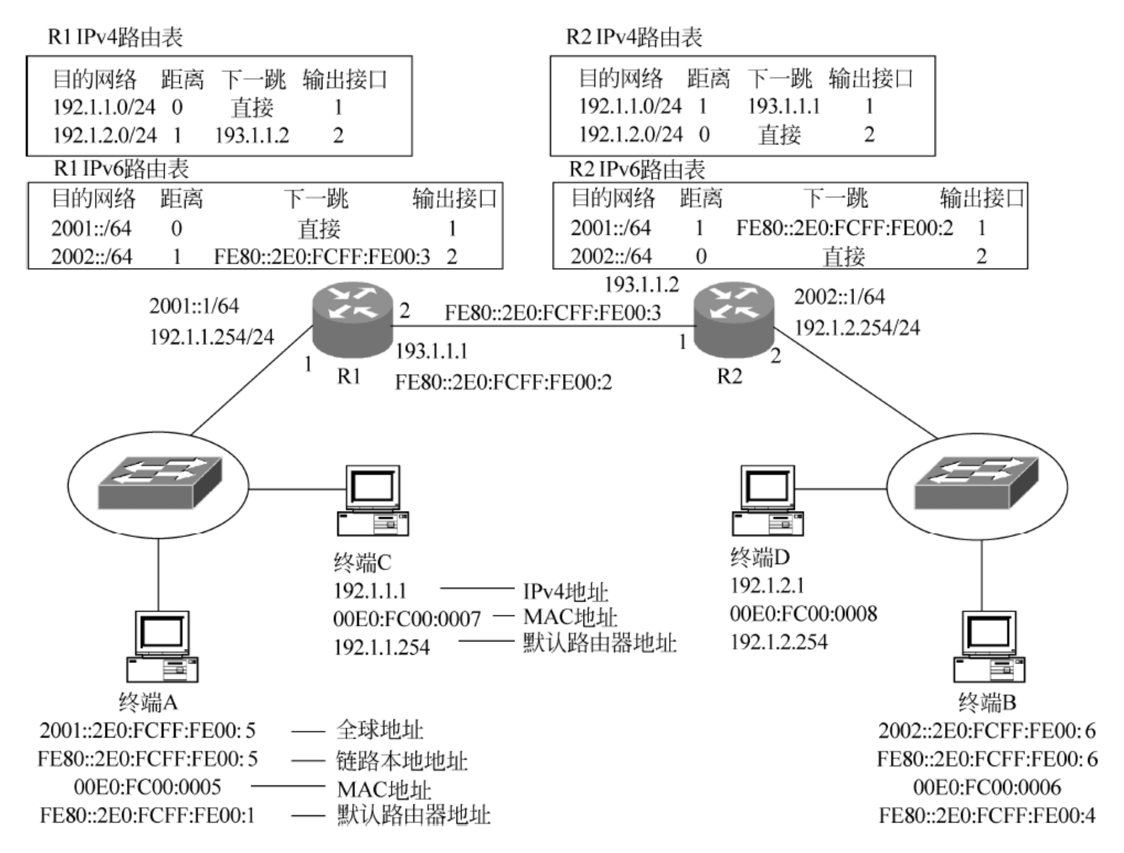


图 10.19 采用双协议栈路由器的网络结构

地址，启动路由协议如 IPv4 的 RIP 和 IPv6 的 RIPng，并通过各自的路由协议建立如图 10.19 所示的 IPv4 和 IPv6 路由表。对于 IPv6 终端，配置相对简单，在采用无状态地址自动配置方式时，自动获取图 10.19 中所示的配置信息。对于 IPv4 终端，通过手工配置，或者通过 DHCP 获取图 10.19 所示的配置信息。无论是终端 A 向终端 B 发送数据，还是终端 C 向终端 D 发送数据，都必须先获取目的终端的 IP 地址，通过目的终端的 IP 地址确定下一跳路由器的 IP 地址，通过 IPv4 over 以太网或 IPv6 over 以太网技术实现下一跳路由器（或目的终端）的地址解析、MAC 帧封装及 MAC 帧传输过程。当路由器接口接收到 MAC 帧，通过 MAC 帧的类型字段确定数据字段包含的 IP 分组类型，将分离出的 IP 分组提交给对应的网络层进程，对应的网络层进程在对应的路由表中完成检索，获取下一跳路由器地址，再次通过 IPv4 over X 或 IPv6 over X（X 指传输网络类型）技术将 IP 分组传输给下一跳路由器，最终将 IP 分组传输给目的终端。需要说明的是，图 10.19 所示的网络结构是无法实现 IPv4 终端和 IPv6 终端之间通信的，除非终端支持双协议栈，否则只能和采用同一网络层协议的终端通信。

10.6.2 隧道技术

双协议栈当然是解决 IPv4 和 IPv6 共存问题的一种有效方法，但当前的 Internet 是 IPv4 网络，路由器只支持 IPv4，而且在短时间内很难使 Internet 中的路由器支持 IPv6，因此，IPv6 网络在未来一段时间内只能是孤岛，无法融入 Internet，图 10.20 给出了 IPv6 网络

的发展路线图。那么,在当前 IPv6 网络为孤岛的情况下,如何实现这些 IPv6 孤岛的互连呢?隧道技术就是一种用于实现 IPv6 孤岛互连的机制。

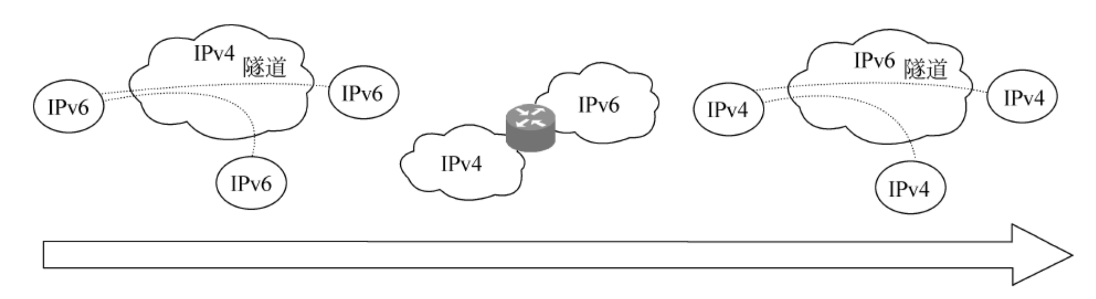


图 10.20 IPv6 网络的发展路线图

图 10.21 是用隧道实现两个 IPv6 孤岛互连的互连网络结构,图中路由器 R1 的接口 2 和路由器 R2 的接口 1 同时配置为 IPv4 和 IPv6 接口,并配置 IPv4 和 IPv6 地址。分别在路由器 R1 和路由器 R2 中定义 IPv4 隧道,隧道两个端点的 IPv4 地址分别为 192.1.1.1 和 192.1.2.2,同时在路由器中设置到达隧道另一端的 IPv4 路由项,路由器配置的信息如图 10.21 所示。当终端 A 需要给终端 B 发送 IPv6 分组时,终端 A 构建以 2001::2E0:FCFF:FE00:5 为源地址,以 2002::2E0:FCFF:FE00:6 为目的地址的 IPv6 分组,并根据配置的默认网关地址将该 IPv6 分组传输给路由器 R1,路由器 R1 用 IPv6 分组的目的地址检索 IPv6 路由表,找到下一跳路由器,但发现连接下一跳路由器的是隧道 1。根据路由器 R1 配置隧道 1 时给出的信息(隧道 1 源地址为 192.1.1.1、目的地址为 192.1.2.2),路由器 R1 将 IPv6 分组封装成隧道格式。由于隧道 1 是 IPv4 隧道,隧道格式外层首部为 IPv4 首部,封装

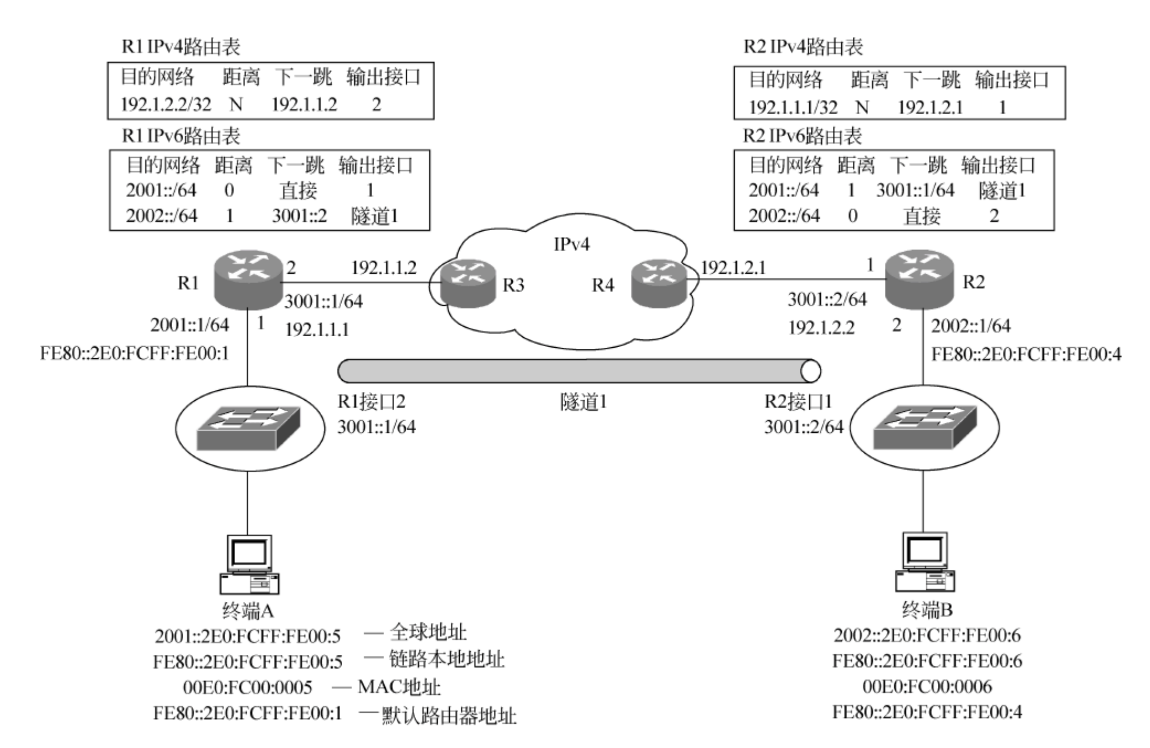


图 10.21 隧道实现两个 IPv6 孤岛互连

后的隧道格式如图 10.22 所示。隧道格式被提交给路由器 R1 的 IPv4 进程，IPv4 进程用隧道格式的目的地址检索 IPv4 路由表，找到下一跳路由器，通过对应的 IPv4 over X 技术将隧道格式转发给路由器 R3。经过 IPv4 网络的逐跳转发，隧道格式到达路由器 R2。由于路由器 R2 的接口 1 被定义成隧道 1 的另一个端点，当路由器 R2 从接口 1 接收到隧道格式，从中分离出 IPv6 分组，并用 IPv6 分组的目的地址检索 IPv6 路由表，找到下一跳结点（目的终端），通过 IPv6 over 以太网技术将 IPv6 分组传输给目的终端。

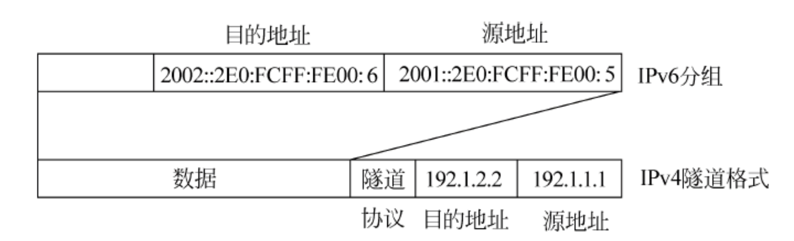


图 10.22 IPv6 分组封装成 IPv4 隧道格式

10.6.3 网络地址和协议转换技术

隧道技术只能解决两个 IPv6 孤岛通过 IPv4 网络进行通信的问题，当 IPv4 网络和 IPv6 网络共存时，更需要一种解决两个分别属于这两种不同网络的终端之间的通信问题的方法，无状态 IP/ICMP 转换（Stateless IP/ICMP Translation，SIIT）与网络地址和协议转换（Network Address Translation-Protocol Translation，NAT-PT）就是解决两个分别属于这两种不同网络的终端之间通信问题的协议。

1. SSIT

当 IPv4 网络中的终端和 IPv6 网络中的终端相互通信时，必须在网络边界实现 IPv4 分组格式和 IPv6 分组格式之间的转换，无状态 IP/ICMP 转换就是一种用于完成 IPv4 分组格式和 IPv6 分组格式之间转换的协议。它需要在 IPv6 网络中为那些需要和 IPv4 网络通信的终端分配 IPv4 地址，但这些 IPv4 地址在 IPv6 网络中被转换成::FFFF:0:a.b.c.d 格式的 IPv6 地址。当 IPv6 网络中的终端希望发送数据给 IPv4 网络中的终端时，它直接在 IPv6 分组的目的地址字段给出 IPv4 网络中的终端的 IPv4 地址，但以::FFFF:a.b.c.d 的 IPv6 地址格式给出。IPv6 网络必须将以::FFFF:a.b.c.d 格式的 IPv6 地址为目的地址的 IPv6 分组路由到网络边界的地址和协议转换器，由地址和协议转换器完成 IPv6 分组格式至 IPv4 分组格式的转换。同样，当 IPv4 网络中的终端希望向 IPv6 网络中的终端发送数据时，它直接在 IPv4 分组的目的地址字段给出分配给 IPv6 网络中终端的 IPv4 地址，IPv4 网络也必须将以分配给 IPv6 网络中终端的 IPv4 地址为目的地址的 IPv4 分组路由到网络边界的地址和协议转换器，由地址和协议转换器完成 IPv4 分组格式至 IPv6 分组格式的转换。下面结合图 10.23 所示的网络结构详细讨论 IPv4 网络中的终端和 IPv6 网络中的终端用 SIIT 实现相互通信的过程。

在图 10.23 所示的网络结构中，分配给 IPv6 网络中终端的 IPv4 网络地址是 193.1.1.0/24，这些地址必须是 IPv4 网络没有使用的地址，IPv4 网络必须保证将目的地址

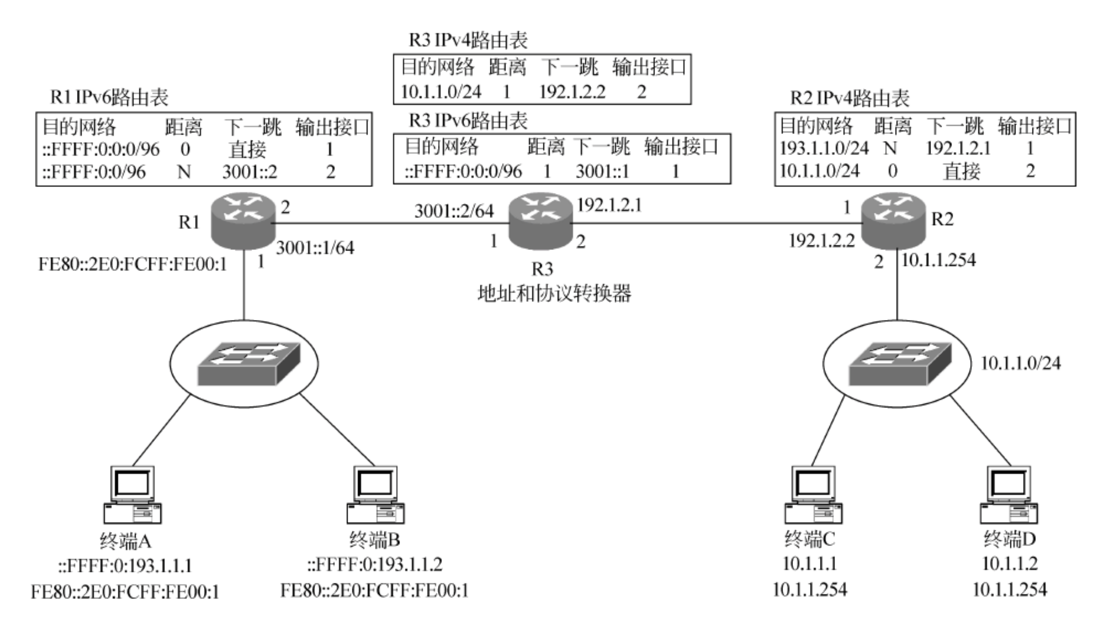


图 10.23 SIIT 实现网络地址和协议转换过程

属于 193.1.1.0/24 的 IPv4 分组路由到 IPv4 网络边界的地址和协议转换器(路由器 R3)，图 10.23 中路由器 R2 路由表中的路由项＜193.1.1.0/24,N,192.1.2.1,1＞就反映了这一点。同样,IPv6 网络也必须保证将以::FFFF:a.b.c.d 格式的 IPv6 地址为目的地址的 IPv6 分组路由到 IPv6 网络边界的地址和协议转换器,图 10.23 中路由器 R1 路由表中的路由项＜::FFFF:0:0/96,N,3001::2,2＞也反映了这一点。作为地址和协议转换器的路由器 R3 支持双协议栈,接口 1 为 IPv6 接口,分配 IPv6 地址。接口 2 为 IPv4 接口,分配 IPv4 地址。通过 IPv6 接口接收到的 IPv6 分组转换成 IPv4 分组后,通过 IPv4 接口转发出去,反之亦然。当图 10.22 中的终端 A 发送数据给终端 C 时,终端 A 构建以::FFFF:0:193.1.1.1 为源地址,::FFFF:10.1.1.1 为目的地址的 IPv6 分组,该 IPv6 分组经过路由器 R1 转发后到达路由器 R3。路由器 R3 完成表 10.6 所示的 IPv6 首部字段至 IPv4 首部字段的转换,用转换后的 IPv4 分组的目的地址(10.1.1.1)检索路由表,找到下一跳路由器,将 IPv4 分组转发给路由器 R2。IPv4 分组经过路由器 R2 转发后,到达终端 C,完成了终端 A 至终端 C 的数据传输过程。

表 10.6 IPv6 首部至 IPv4 首部转换

IPv6 首部字段	IPv4 首部字段
版本：6	版本：4
	首部长度：5
信息流类别：X	服务类型：X
净荷长度：Y	总长度：Y+20(20 是 IPv4 首部长度)
	标识：0
	MF=0,DF=1
	片偏移：0
跳数限制：Z	生存时间：Z
下一个首部：A	协议：A
	首部检验和：重新计算

续表

IPv6 首部字段	IPv4 首部字段
源地址：::FFFF:0:193.1.1.1 目的地址：::FFFF:10.1.1.1	源地址：193.1.1.1 目的地址：10.1.1.1

当终端 C 向终端 A 发送数据时，终端 C 构建以 10.1.1.1 为源地址，193.1.1.1 为目的地址的 IPv4 分组，该 IPv4 分组经过路由器 R2 转发后到达路由器 R3，路由器 R3 完成表 10.7 所示的 IPv4 首部字段至 IPv6 首部字段的转换，用转换后的 IPv6 分组的目的地址检索路由表，找到下一跳路由器，将 IPv6 分组转发给路由器 R1。IPv6 分组经过路由器 R1 转发后，到达终端 A，完成了终端 C 至终端 A 的数据传输过程。为简单起见，假定 IPv4 分组和 IPv6 分组都没有任何可选项或扩展首部，其格式如图 10.24 所示。

表 10.7　IPv4 首部至 IPv6 首部转换

IPv4 首部字段	IPv6 首部字段
版本：4 服务类型：X 总长度：Y 协议：A 生存时间：Z 源地址：10.1.1.1 目的地址：193.1.1.1	版本：6 信息流类别：X 流标签：0 净荷长度：Y－20(20 是 IPv4 首部长度) 下一个首部：A 跳数限制：Z 源地址：::FFFF:10.1.1.1 目的地址：::FFFF:0:193.1.1.1

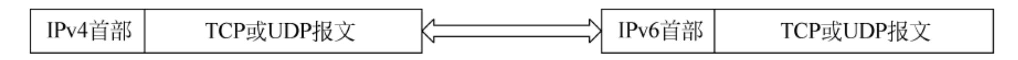


图 10.24　IPv4 和 IPv6 分组格式

IPv6 分组转换成 IPv4 分组时，一些 IPv4 首部字段值可以直接从对应的 IPv6 首部字段中复制过来，如服务类型、生存时间、协议。一些 IPv4 首部字段值可以通过对应的 IPv6 首部字段值导出，如总长度、源地址、目的地址。一些 IPv4 首部字段值只能设置成约定值，如标识、片偏移、MF 和 DF 标志位。

同样，IPv4 分组转换成 IPv6 分组时，一些 IPv6 首部字段值可以直接从对应的 IPv4 首部字段中复制过来，如信息流类别、下一个首部、跳数限制。一些 IPv6 首部字段值可以通过对应的 IPv4 首部字段值导出，如净荷长度、源地址、目的地址。一些 IPv6 首部字段值只能设置成约定值，如流标签。

SIIT 能够比较简单地解决属于两种不同网络(IPv4 和 IPv6 网络)的终端之间通信问题，但需要为 IPv6 网络中的终端分配 IPv4 地址，而且这种地址分配是静态的，IPv6 网络中只有分配了 IPv4 地址的终端才能和 IPv4 网络中的终端通信。这就可能需要为 IPv6 网络分配大量的 IPv4 地址，而引发 IPv6 的最主要原因就是 IPv4 地址短缺问题，因此，这种通过对 IPv6 网络中的终端静态分配 IPv4 地址来解决 IPv4 终端和 IPv6 终端之间通信问题的方式存在很大的局限性。多数情况下，虽然 IPv6 网络中有多个终端需要和 IPv4 网络中的终

端通信,但需要同时通信的终端并不多,因此,可以只对 IPv6 网络分配少许 IPv4 地址,以此构成 IPv4 地址池,IPv6 网络中需要和 IPv4 网络通信的终端临时从地址池中分配一个空闲的 IPv4 地址,在通信结束后自动释放该 IPv4 地址。由于每一个 IPv4 地址都不固定分配给 IPv6 网络中的终端,将这种地址分配方式称为动态地址分配方式,而第 8 章讨论的动态 NAT 就是这样一种分配机制。

2. NAT-PT

1) 单向会话通信过程

网络地址和协议转换(Network Address Translation-Protocol Translation,NAT-PT)是一种将 SIIT 和动态 NAT 有机结合的地址和协议转换技术,它对 IPv6 网络中终端的地址配置没有限制,也不需要对 IPv6 网络中需要和 IPv4 网络通信的终端分配 IPv4 地址。它和 IPv4 网络所采用的动态 NAT 一样,在网络边界的地址和协议转换器设置一组 IPv4 地址,并以此构成 IPv4 地址池,当 IPv6 网络中的某个终端发起和 IPv4 网络中的终端之间的会话时,由地址和协议转换器为发起会话的终端分配一个 IPv4 地址,并将该 IPv4 地址和该终端发起的会话绑定在一起。如果会话是 TCP 连接,则可用会话两端的源地址和目的地址、源端口号和目的端口号来标识该会话。在会话存在期间,该 IPv4 地址一直分配给发起会话的终端,当属于该会话的 IPv6 分组经过地址和协议转换器进入 IPv4 网络时,用该 IPv4 地址取代 IPv6 分组的源地址,并完成 IPv6 分组至 IPv4 分组的转换。IPv4 网络中的终端用该 IPv4 地址和发起会话的终端通信,当属于该会话的 IPv4 分组进入地址和协议转换器时,用该 IPv4 分组的目的地址检索会话表,用会话表中给出的发起会话终端的 IPv6 地址取代 IPv4 分组的目的地址,并完成 IPv4 分组至 IPv6 分组的转换。在 SIIT 中,IPv6 网络用::FFFF:a.b.c.d 格式表示 IPv4 地址 a.b.c.d,在 NAT-PT 中,96 位网络前缀可以是其他的值,但必须保证 IPv6 网络将目的地址和该 96 位网络前缀匹配的 IPv6 分组路由到网络边界的地址和协议转换器。地址和协议转换器将和 96 位网络前缀匹配的目的地址的低 32 位作为 IPv4 地址。反之,地址和协议转换器在 IPv4 分组的源地址前加上 96 位网络前缀后作为 IPv6 分组的源地址。下面结合图 10.25 详细讨论一下 NAT-PT 的工作机制。

在图 10.25 中,当终端 A 需要向终端 C 传输数据时,终端 A 发送一个以 2001::2E0:FCFF:FE00:7 为源地址,以 2::10.1.1.1 为目的地址的 IPv6 分组,该 IPv6 分组被 IPv6 网络路由到路由器 R3。路由器 R3 用该 IPv6 分组的源地址检索地址转换表,由于这是终端 A 发送给 IPv4 网络的第一个 IPv6 分组,地址转换表不存在匹配的地址转换项,路由器 R3 为终端 A 分配一个 IPv4 地址,这里假定是 193.1.1.1,同时,在地址转换表中创建一项用于建立 IPv6 地址 2001::2E0:FCFF:FE00:7 与 IPv4 地址 193.1.1.1 之间映射的地址转换项,如表 10.8 所示。路由器 R3 将该 IPv6 分组转换成 IPv4 分组,通过 IPv4 路由表确定的传输路径将该 IPv4 分组转发给下一跳路由器 R2。该 IPv4 分组经过路由器 R2 转发后到达终端 C,完成终端 A 至终端 C 的传输过程。IPv6 分组转换成 IPv4 分组时各字段的转换过程和 SIIT 相同,如表 10.6 所示。源和目的地址的转换过程如图 10.26 所示。

表 10.8 地址转换表

IPv6 地址	IPv4 地址
2001::2E0:FCFF:FE00:7	193.1.1.1

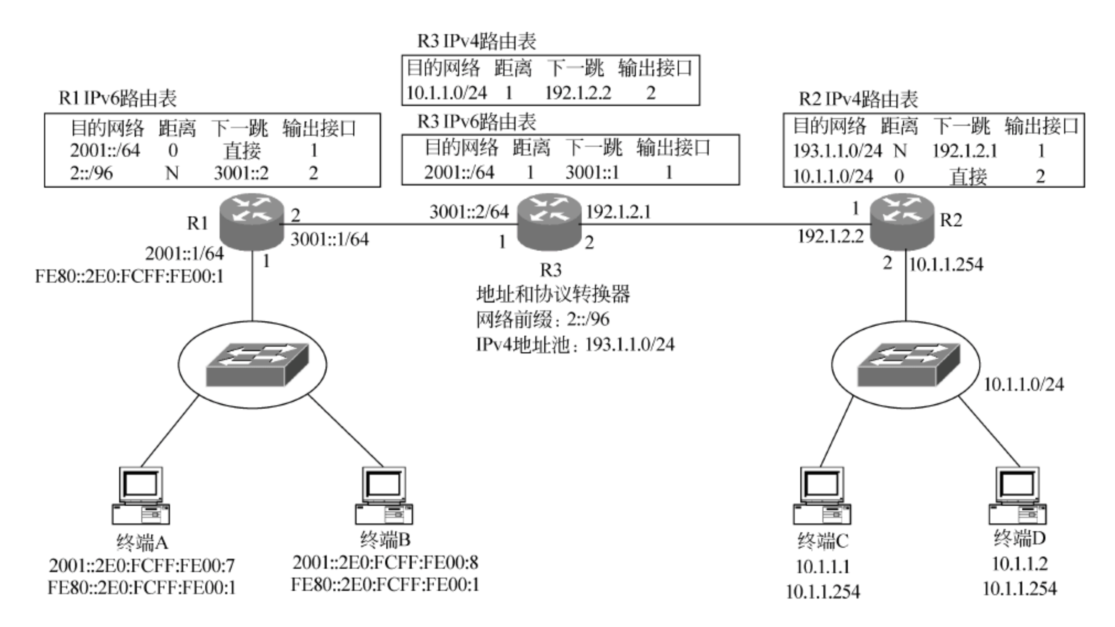


图 10.25 NAT-PT 实现网络地址和协议转换过程

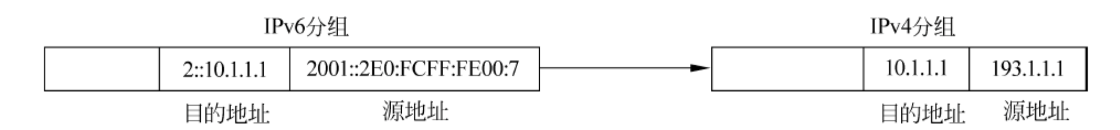


图 10.26 IPv6 分组至 IPv4 分组转换过程

当终端 C 需要向终端 A 发送数据时，终端 C 构建一个以 10.1.1.1 为源地址，以 193.1.1.1为目的地址的 IPv4 分组，该 IPv4 分组被 IPv4 网络路由到路由器 R3。路由器 R3 用该 IPv4 分组的目的地址检索地址转换表，找到匹配的地址转换项，用该地址转换项中的 IPv6 地址作为转换后的 IPv6 分组的目的地址。由于为路由器 R3 配置的网络前缀为 2::/96，源地址被转换成 2::10.1.1.1。IPv4 分组转换成 IPv6 分组时各字段的转换过程和 SIIT 相同，如表 10.7 所示。源和目的地址的转换过程如图 10.27 所示。

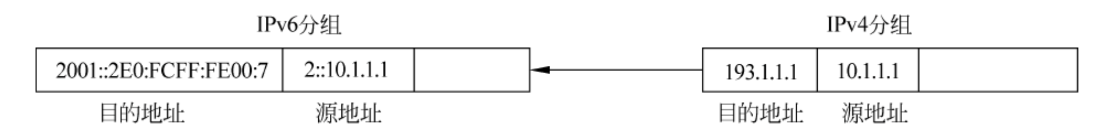


图 10.27 IPv4 分组至 IPv6 分组转换过程

终端 A 后续发送给终端 C 的 IPv6 分组，由于能够在地址转换表中找到匹配的地址转换项，可以根据该地址转换项中的 IPv4 地址进行源地址转换。地址转换表中的每一项地址转换项都关联一个定时器，每当通过路由器 R3 连接 IPv6 网络的接口接收到源地址为该地址转换项中 IPv6 地址的 IP 分组，刷新与该地址转换项关联的定时器，一旦关联的定时器溢出，将删除该地址转换项，路由器可以重新分配该地址转换项中的 IPv4 地址。

2）双向会话通信过程

和 IPv4 动态 NAT 一样，NAT-PT 只能用于由 IPv6 网络中的终端发起会话的应用，如果某个应用需要由 IPv4 网络中的终端发起会话，NAT-PT 是无法实现的，因为，IPv4 网络

中的终端是无法用某个 IPv4 地址来绑定 IPv6 网络中的某个终端的。如果非要实现由 IPv4 网络中的终端发起的会话，需要采用静态 NAT，即在路由器 R3 配置静态的 IPv4 地址和 IPv6 地址之间的映射。如图 10.28 中，如果终端 C 希望访问 IPv6 网络中的 DNS 服务器(IPv6 DNS)，就构建以 10.1.1.1 为源地址，以 193.1.1.5 为目的地址的 IPv4 分组，该 IPv4 分组到达路由器 R3 后，路由器 R3 通过手工配置的静态地址映射，将目的地址转换成 2001::2E0:FCFF:FE00:9。但如果对 IPv6 中的其他终端也采用静态地址映射，NAT-PT 将重新变为 SIIT，需要为所有可能和 IPv4 网络通信的 IPv6 网络中的终端静态分配 IPv4 地址，这显然是不可能的。对于图 10.28 所示的网络结构，路由器 R3 不仅是地址和协议转换器，还是 DNS 应用层网关，DNS 用于将完全合格的域名解析成 IP 地址，如果是 IPv6 网络，则解析成 IPv6 地址，如果是 IPv4 网络，则解析成 IPv4 地址。DNS 服务器给出完全合格的域名和对应的 IP 地址之间的映射，如＜终端 A：2001::2E0:FCFF:FE00:7＞。DNS 应用层网关完成 IPv4 DNS 协议和 IPv6 DNS 协议之间的转换，这种转换除了消息格式转换外，还包括命令和响应的转换。当终端 C 想发起和终端 A 之间的会话时，首先通过 DNS 解析出终端 A 的完全合格的域名——终端 A 所对应的 IPv4 地址。由于在路由器 R3 中已经静态配置了 IPv6 网络中 DNS 服务器的 IPv6 地址 2001::2E0:FCFF:FE00:9 和 IPv4 地址 193.1.1.5 之间的映射，终端 C 配置的 DNS 服务器地址为 193.1.1.5，因此，当需要 DNS 解析出完全合格的域名——终端 A 所对应的 IPv4 地址时，向 IPv4 地址为 193.1.1.5 的 DNS 服务器发送请求报文，请求报文被封装成 IPv4 分组后进入 IPv4 网络，被 IPv4 网络路由到路由器 R3。由路由器 R3 完成 IPv4 DNS 请求报文至 IPv6 DNS 请求报文的转换，并将该 DNS 请求报文封装成以 2::10.1.1.1 为源地址，以 2001::2E0:FCFF:FE00:9 为目的地址的 IPv6 分组，通过 IPv6 网络将该 IPv6 分组传输到 IPv6 网络的 DNS 服务器。IPv6 网络的 DNS 服务器根据完全合格的域名——终端 A 解析出 IPv6 地址——2001::2E0:FCFF:FE00:7，并将该地址通过 DNS 响应报文回送给地址为 2::10.1.1.1 的终端(终端 C)。该 DNS 响应报文被 IPv6 网络路由到路由器 R3，由路由器 R3 在 IPv4 地址池中选择一个未分配的 IPv4 地址，这里假定是 193.1.1.1，将其分配给终端 A，同时在地址转换表创建用于建立 2001::2E0:FCFF:FE00:7 和 193.1.1.1 之间映射的地址转换项。路由器 R3 将 IPv6 DNS 响应报文转换为 IPv4 DNS 响应报文，并将该 IPv4 DNS 响应报文封装成以 10.1.1.1 为目的地址的 IPv4 分组，通过 IPv4 网络将该 IPv4 分组传输到终端 C，终端 C 随后用 IPv4 地址193.1.1.1和终端 A 进行通信。需要指出的是，在上述通信过程中，IPv4 网络中的终端通过 DNS 的地址解析过程创建用于建立 2001::2E0:FCFF:FE00:7 和 193.1.1.1 之间映射的地址转换项，路由器 R3 将所有通过连接 IPv6 网络接口接收到的源地址为 2001::2E0:FCFF:FE00:7 的 IPv6 分组转换成源 IP 地址为 193.1.1.1 的 IPv4 分组，将所有通过连接 IPv4 网络接收到的目的地址为 193.1.1.1 的 IPv4 分组转换成目的地址为 2001::2E0:FCFF:FE00:7 的 IPv6 分组。通过 DNS 的地址解析过程创建的地址转换项等同于动态 NAT 创建的地址转换项。

IPv4 网络中所有终端和服务器对应的 IPv6 地址是固定的，IPv6 网络中的终端可以获取 IPv4 网络中所有终端和服务器对应的 IPv6 地址，因此，IPv6 网络中的终端可以通过直接给出 IPv6 地址的方式和 IPv4 网络中的终端通信。当然，记住完全合格的域名总比记住 128 位的 IPv6 地址容易，因此，IPv6 网络中的终端可能通过完全合格的域名(如终端 C)发

起和 IPv4 网络中的终端之间的会话。这种情况下，由 IPv6 终端向 IPv4 网络的 DNS 服务器发送 DNS 请求报文，由路由器 R3 完成 IPv6 DNS 请求报文至 IPv4 DNS 请求报文的转换。当路由器 R3 接收到 IPv4 网络中的 DNS 服务器回送的 DNS 响应报文时，一方面通过加上网络前缀 2::，将解析出的 IPv4 地址转换成 IPv6 地址；另一方面完成 IPv4 DNS 响应报文至 IPv6 DNS 响应报文的转换。

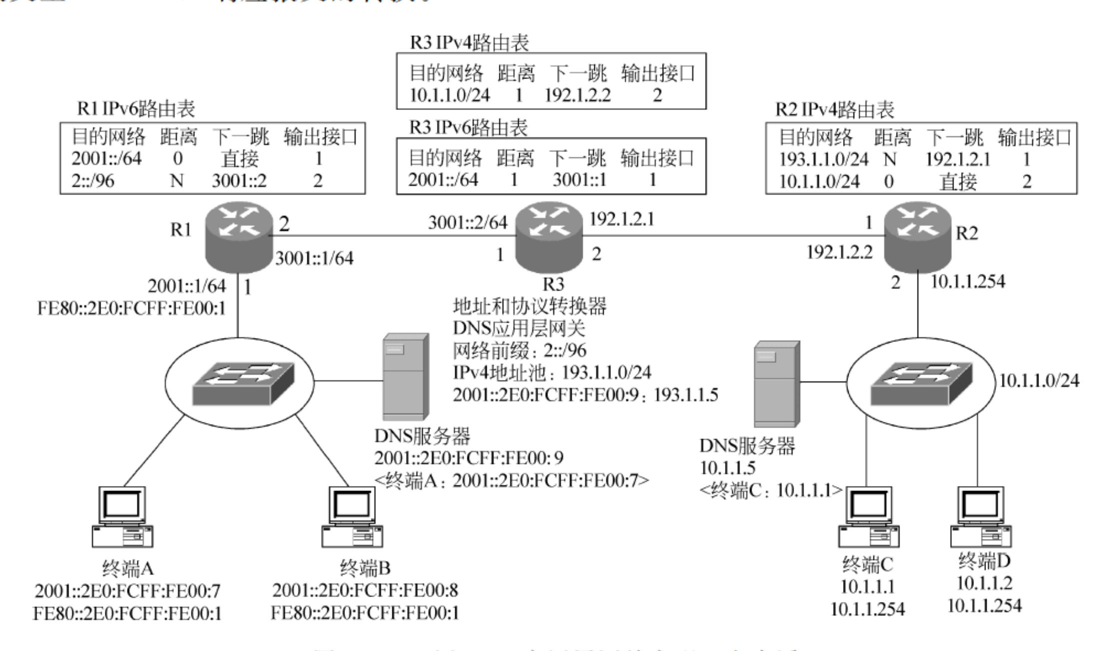


图 10.28 用 DNS 应用层网关实现双向会话

隧道技术和 NAT-PT 都是解决 IPv6 网络和 IPv4 网络互连的权宜之计，也都存在很多问题，有很大的局限性，它们只能解决 IPv4 或 IPv6 网络中一方占据主导地位时，和作为孤岛的另一方终端的通信问题。因此，如果 IPv6 网络和 IPv4 网络长时间平分秋色的话，必须用更合适的技术来解决它们的互连问题。目前情况下，IPv4 占据绝对的主导地位，但网络发展的趋势是用 IPv6 代替 IPv4，只是不知道这个过程何时开始，需要多长时间。

习题

10.1 IPv4 的主要缺陷有哪些？

10.2 IPv4 短时间内是否会被 IPv6 取代？并解释为什么。

10.3 IPv6 和 IPv4 相比，有什么优势？

10.4 这样设计 IPv6 首部的理由是什么？增加的字段有什么作用？

10.5 IPv6 取消首部检验和字段的理由是什么？

10.6 IPv6 的扩展首部是否只是取代 IPv4 的可选项？它有什么作用？

10.7 IPv6 分片过程和 IPv4 分片过程相比，有哪些优势？

10.8 IPv6 地址结构的设计依据是什么？

10.9 将以下用基本表示方式表示的 IPv6 地址用零压缩表示方式表示。

(1) 0000:0000:0F53:6382:AB00:67DB:BB27:7332。

(2) 0000:0000:0000:0000:0000:0000:004D:ABCD。

(3) 0000:0000:0000:AF36:7328:0000:87AA:0398。

(4) 2819:00AF:0000:0000:0000:0035:0CB2:B271。

10.10 将以下用零压缩表示方式表示的IPv6地址用基本表示方式表示。

(1) ::。

(2) 0:AA::0。

(3) 0:1234::3。

(4) 123::1:2。

10.11 给出以下每一个IPv6地址所属的类型。

(1) FE80::12。

(2) FEC0::24A2。

(3) FF02::0。

(4) 0::01。

10.12 下述地址表示方法是否正确。

(1) ::0F53:6382:AB00:67DB:BB27:7332。

(2) 7803:42F2: ::88EC:D4BA:B75D:11CD。

(3) ::4BA8:95CC::DB97:4EAB。

(4) 74DC::02BA。

(5) ::00FF:128.112.92.116。

10.13 IPv6为什么没有广播地址？哪个组播地址等同于全1的广播地址？

10.14 IPv6设置链路本地地址的目的是什么？

10.15 为什么使用无状态地址自动配置方式？IPv4为什么不使用这种地址分配方式？

10.16 IPv4是否不需要重复地址检测？如果需要，如何实现重复地址检测？

10.17 分别用IPv6和IPv4设计一个有30个终端的交换式以太网，并使各个以太网内的终端之间能够相互通信，给出设计步骤，并比较其过程。

10.18 IPv4 over以太网用ARP实现目的终端地址解析，ARP报文直接用MAC帧封装，而IPv6 over以太网用邻站发现协议实现目的终端地址解析，用IPv6分组封装邻站发现协议的协议报文，这两者有什么区别？

10.19 根据图10.29所示的网络结构，配置终端和三层交换机，并讨论终端A至终端B的IPv6分组传输过程。

10.20 根据图10.30所示的网络结构，配置终端和三层交换机，讨论三层交换机之间用RIPng建立路由表过程，并给出终端A至终端D的IPv6分组传输过程。

10.21 IPv4和IPv6互连的技术有哪些？各自在什么应用环境下使用？

10.22 假定图10.30中，VLAN 2使用IPv4，其他VLAN使用IPv6，请给出用双协议栈解决IPv4和IPv6网络共存和同一网络内终端之间通信问题的配置，并讨论终端B至终端C、终端A至终端D之间的通信过程。

10.23 假定图10.30中，VLAN 3使用IPv4，其他VLAN使用IPv6，请给出用SIIT解决属于不同类型网络的终端之间通信问题的配置，并讨论终端A至终端B、终端C至终

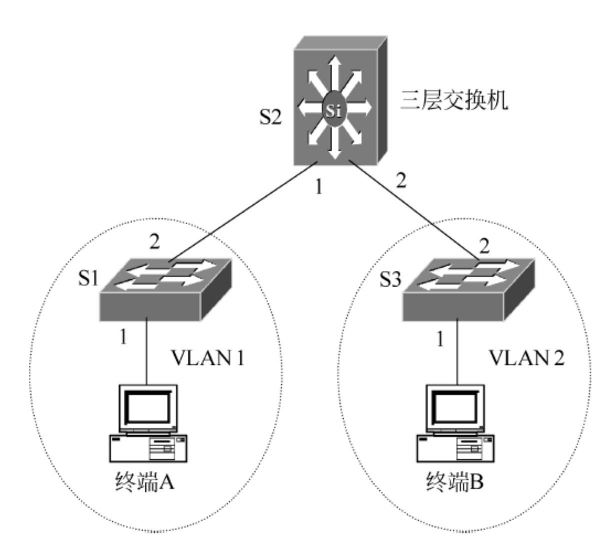


图 10.29 题 10.19 图

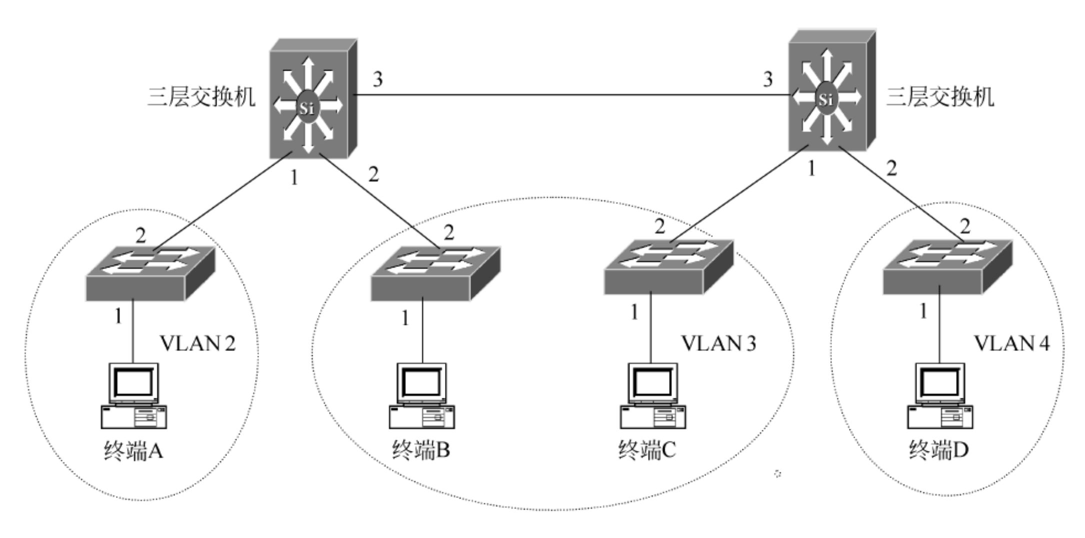


图 10.30 题 10.20 图

端 D 之间的通信过程。

10.24 假定图 10.30 中,VLAN 3 使用 IPv4,其他 VLAN 使用 IPv6,请给出用 NAT-PT 和 DNS 应用层网关解决属于不同网络的终端之间通信问题的配置,并讨论终端 A 至终端 B、终端 C 至终端 D 之间的通信过程。

10.25 SIIT 的局限性是什么?

10.26 NAT-PT 的局限性是什么?

10.27 NAT-PT 实现双向会话的原理是什么?

10.28 能否仿照 IP 互连不同类型传输网络的模式,提出一种真正实现 IPv4 和 IPv6 网络互连的模式?

英文缩写词

ALG(Application Level Gateway)应用层网关(8.2)
ARP(Address Resolution Protocol)地址解析协议(5.3)
AS(Autonomous System)自治系统(6.2)
ASBR(Autonomous System Boundary Router)自治系统边界路由器(6.2)
ATM(Asynchronous Transfer Mode)异步传输模式(1.2)
BDR(Backup Designated Router)备份指定路由器(6.4)
BGP(Border Gateway Protocol)边界网关协议(6.2)
BPDU(Bridge Protocol Data Unit)网桥协议数据单元(3.2)
BR(Bootstrap Router)引导路由器(7.3)
CAM(Content Addressable Memory)内容寻址存储器(1.3)
CIDR(Classless InterDomain Routing)无分类编址(5.2)
CIST(Common and Internal Spanning Tree)公共内部生成树(3.4)
CRC(Cyclic Redundancy Check)循环冗余检验(1.2)
CSMA/CD(Carrier Sense Multiple Access/ Collision Detection)载波侦听多点接入/冲突检测 (1.2)
CST(Common Spanning Tree)公共生成树(3.4)
DAD(Duplicate Address Detection)重复地址检测(10.4)
DD(Database Description)数据库描述(6.4)
DiffServ(Differentiated Services)区分服务(10.2)
DNS(Domain Name System)域名系统(8.1)
DR(Designated Router)指定路由器(6.4)
DS(Differentiated Services)区分服务(10.2)
DSCP(Differentiated Services Code Point)区分服务码点(10.2)
DVMRP(Distance Vector Multicast Routing Protocol)距离向量组播路由协议(7.1)
EGP(External Gateway Protocol)外部网关协议(6.2)
FCS(Frame Check Sequence)帧检验序列(1.2)
FDDI(Fiber Distributed Data Interface)光纤分布式数据接口(1.1)
FTP(File Transfer Protocol)文件传输协议(8.1)
GARP(Generic Attribute Registration Protocol)通用属性注册协议(2.6)
GVRP(GARP VLAN Registration Protocol)VLAN 属性注册协议(2.6)
IFG(Inter Frame Gap)最小帧间间隔(1.2)
IGMP(Internet Group Management Protocol)互联网组管理协议(7.1)
IGP(Interior Gateway Protocol)内部网关协议(6.2)

IntServ(Integrated Services) 综合服务(10.2)
IP(Internet Protocol)网际协议(5.1)
ISP(Internet Service Provider)Internet 服务提供者(8.1)
IST(Internal Spanning Tree)内部生成树(3.4)
LACP(Link Aggregation Control Protocol)链路聚合控制协议(4.1)
LAN(Local Area Network)局域网(1.1)
LLC(Logical Link Control)逻辑链路控制(1.1)
LSA(Link State Advertisement)链路状态通告(6.4)
LSR(Link State Request)链路状态请求(6.4)
LSU(Link State Update)链路状态更新(6.4)
MAC(Medium Access Control)媒体接入控制(1.1)
MADCAP(Multicast Address Dynamic Client Allocation Protocol)组播地址动态客户端分配协议(7.1)
MAN(Metropolitan Area Network)城域网(1.4)
MSTI(Multiple Spanning Tree Instance)多生成树实例(3.4)
MSTP(Multiple Spanning Tree Protocol)多生成树协议(3.1)
MTU(Maximum Transfer Unit)最大传输单元(5.2)
NAT(Network address translation)网络地址转换(8.1)
NAT-PT(Network Address Translation-Protocol Translation)网络地址和协议转换(10.6)
ND(Neighbor Discovery)邻站发现(10.4)
NIC(Network Interface Card)网络接口卡(1.1)
OSI/RM(Open Systems Interconnection/Reference Model)开放系统互连/参考模型(1.2)
OSPF(Open Shortest Path First)开放最短路径优先(6.2)
PAT(Port Address Translation)端口地址转换(8.2)
PDA(Personal Digital Assistant)个人数字助理(10.1)
PDU(Protocol Data Unit)协议数据单元(8.1)
PIM-DM(Protocol Independent Multicast-Dense Mode)协议无关组播—密集方式(7.1)
PIM-SM(Protocol Independent Multicast-Sparse Mode)协议无关组播—稀疏方式(7.1)
PPP(Point-to-Point Protocol)点对点协议(5.1)
PSTN(Public Switched Telephone Network)公共交换电话网(5.1)
RIP(Routing Information Protocol)路由信息协议(6.2)
RIPng(RIP Next Generation)下一代 RIP()
RP(Rendezvous Point)汇聚点(7.3)
RSTP(Rapid Spanning Tree Protocol)快速生成树协议(3.1)
SDH(Synchronous Digital Hierarchy)同步数字体系(1.4)
SDT(Session Directory Tool)会话目录工具(7.1)
SIIT(Stateless IP/ICMP Translation)无状态 IP/ICMP 转换(10.6)
STP(Spanning Tree Protocol)生成树协议(1.1)
TC(Topology Change)拓扑改变(3.2)

TCA(Topology Change Acknowledgment)拓扑改变应答(3.2)
TCN(Topology Change Notification) 拓扑改变通知(3.2)
VLAN(Virtual LAN)虚拟局域网(2.1)
VMPS(VLAN Membership Policy Server)VLAN 成员策略服务器(2.4)
VPN 虚拟专用网络(Virtual Private Network)(8.1)
VRID(Virtual Router IDentifier)虚拟路由器标识符(5.4)
VRRP(Virtual Router Redundancy Protocol)虚拟路由器冗余协议(5.4)
VTP(VLAN Trunking Protocol)VLAN 主干协议(2.6)

参 考 文 献

1. Larry L. Peterson,Bruce S. Davie. Computer Networks,A Systems Approach Fourth Edition. 北京：机械工业出版社,2008.
2. Andrew S. Tanenbaum. Computer Networks Fourth Edition. 北京：清华大学出版社,2004.
3. Kennedy Clark,Kevin Hamilton. Cisco LAN Switching. 北京：人民邮电出版社,2003.
4. Jeff Doyle 著. TCP/IP 路由技术(第一卷). 葛建立,吴剑章译. 北京：人民邮电出版社,2003.
5. Jeff Doyle,Jennifer DeHaven Carroll. TCP/IP 路由技术(第二卷). 北京：人民邮电出版社,2003.
6. 谢希仁. 计算机网络(第 5 版). 北京：电子工业出版社,2009.
7. 沈鑫剡等. 计算机网络技术及应用. 北京：清华大学出版社,2007.
8. 沈鑫剡. 计算机网络. 北京：清华大学出版社,2008.
9. 沈鑫剡等. 计算机网络技术及应用(第 2 版). 北京：清华大学出版社,2010.
10. 沈鑫剡. 计算机网络(第 2 版). 北京：清华大学出版社,2010.
11. 沈鑫剡等. 计算机网络技术及应用学习辅导和实验指南. 北京：清华大学出版社,2011.
12. 沈鑫剡,叶寒锋. 计算机网络学习辅导与实验指南. 北京：清华大学出版社,2011.